Statistical Methods to Incorporate External Summary-level Information into a Current Study

by

Tian Gu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2021

Doctoral Committee:

Professor Bhramar Mukherjee, Co-Chair
Professor Jeremy M.G. Taylor, Co-Chair
Assistant Professor Yang Chen
Assistant Professor Peisong Han

Tian Gu
gutian@umich.edu
ORCID iD: 0000-0002-0856-8843

ACKNOWLEDGMENTS

It is unbelievable that I am coming to the end of my graduate school journey. I have had an enjoyable and rewarding six years in Ann Arbor, and I would like to take this opportunity to express my sincere gratitude to many people.

Thanks go first to my Ph.D. co-advisors, Drs. Jeremy M.G. Taylor and Bhramar Mukherjee. I have enjoyed and looked forward to every meeting we had in the past 4 years, in-person or virtually, which happened nearly every week since 2017, rain or shine, even when Dr. Taylor was in New Zealand for a sabbatical or I was in China over the winter breaks. Before the pandemic, I would drop by at Dr. Taylor's office for an extra "office hour" on Friday afternoon, while Bhramar would sometimes invite me to the tea shop for a Sunday afternoon chat, just like friends. In addition to the undoubted A^{+} mentoring that I experienced from both of them, I also learned from their unique and charming personalities. Dr. Taylor has been very responsive to all my emails and responsible for every single sentence I wrote in the manuscript. He showed me the most thorough and meticulous attitude, rigorous reasoning, and clear writing in revision after revision until the work met his satisfaction. Bhramar has been brimful of energy and enthusiasm for work, life, and art. Seeing her take on new challenges, balance different roles, accommodate multi-relationships in work, and yet still produce solid work and keep a high spirit in life has influenced me positively. Their knowledgeable minds, humble and curious attitude to research, and passion for novelty provided me support and guidance of dissertation mentorship and how to be a responsible researcher. They will always have my deepest admiration and respect.

My sincere appreciation also goes to Dr. Susan Murray, my advisor for my graduate student research assistant position. Susan has been patiently teaching me every detail in work, turning all tasks into learning opportunities, and recognizing every credit I earned. I truly benefited from her critical thinking in applied statistical analysis and excellent communication skills with non-statisticians. Susan was never sparing with praise and encouragement, and she showed exceptional support and respect towards all my choices. It has been a delightful and rewarding experience working with her and the wonderful collaborators at the pulmonary group in the University of Michigan Medical School: Meilan Han, Wassim Labaki, Margaret Salisbury, Beth Belloli, and Vebha Lama.

Many thanks to members of my committee, Drs. Peisong Han and Yang Chen. Peisong has inspired me to solve the largest obstacle in my second project, a derivation that troubled me for over a half year. He kindly endured my many unannounced visits. Yang has been supportive and responsive to all my requests. I am honored to have had them as my committee members.

In addition to my co-advisors and committee members, I would like to thank professors from the data integration working group–Drs. Vera Baladandayuthapani, Peter Song, Phil Boonstra, Kevin He, Jian Kang, and Hui Jiang, who listened to many of my immature presentations about my preliminary research results and provided constructive insights; especially Dr. Hui Jiang, who guided me on my very first research project in graduate school over the summer of 2016.

I want to thank my lab-mates in both Mukherjee and TaBaBoo lab, especially Dr. Lauren Beesley, who is a master at balancing work and life, and a great friend and mentor who always inspires me academically and mentally; Max Salvatore and Dr. Lars Fritsche, who were the best teammate and mentor I could ask for in the COVID projects; Dr. Phil Boonstra, my academic brother who was kind enough to spent time looking at my R code in my first project and teach me many great R tricks; and Emily Roberts, Elizabeth Chase, Jiacong Du, and Pedro Orozco Del Pino, who have excellent personalities and have been great friends of mine.

I am lucky to have many classmates and friends who supported each other along the way. My graduate school experience would not have been so colorful without the following lovely people: Youfei Yu, Nina Zhou, Ming Tang, Yingchao Zhong, Yiwang Zhou, Zhangcheng Zhao, Jingyi Zhai, Yatong Li, Lan Luo, Jiaqiang Zhu, Yuqing Lin, Boran Gao, and Yancheng Zhao. In addition, I want to thank our school of public health family, the faculty and staff that have made Michigan Biostats feel like home, especially Nicole Fenech, Irene Felicetti, Dr. Kirsten Herold and Dr. Xu Shi.

Shout out to my cutest current office-mates who form the best office in SPH–Lulu Shang, Fatema Shafie, Kim Hochstedler, Tahmeed Tureen, and Jung Yeon Won. Some of my old office-mates and senior peers showed me how to be a Ph.D. student and explore career paths in my early years, including Drs. Jinchunzi Shi, Yumeng Li, Sheng Qiu, Summer Xia, Boxian Wei, Marco Benedetti, Chris Lee and Vincent Tan. Outside school, the friends I made in the Chinese rowing team–Jiayi Tian, Lu Qu, Ting Lin, Tiancheng Zuo, and Yuhan Liang have witnessed a precious time with me between 2018 and 2020.

I want to show my special gratitude to one of my oldest friends, Lv Ni, who has been on my side since middle school and flew from Wisconsin to participate in my master's graduation ceremony. She has always been generous with her time to discuss any questions that I had in my research, even though her schedule is packed with teaching, research, and tedious administrative work as a young assistant professor. She is my go-to person whenever I am stuck in a math problem. I feel safe to have her backing, and she absolutely deserves an independent paragraph in these

acknowledgments.

Finally, I want to thank my parents, Ruijun Wu and Renxu Gu, and my boyfriend, Zizheng Zhang. My parents have always been my role models, giving me so much trust, freedom, and unconditional support. I admire my mom's determination and hard work as a researcher and her courage and vision as a female leader. As a strict father, my dad helped me develop good habits in living and studying when I was little. Although he participated less in my academic life when my college major was completely out of his specialty, he has always been the cheerleader for my mom and me, and he always takes good care of us. Finally, Zizheng has contributed a lot in helping me manage my stress. Although Zizheng has been the biggest distraction of my studying, without his company, I could not have survived the 451-day-long COVID-quarantine working from home between March 14, 2020, and June 8, 2021.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

LIST OF ACRONYMS

ARE asymptotic relative efficiency

AUC area under the curve

BS scaled Brier score

CML constrainted maximum likelihood

CSPML constrained semi-parametric maximum likelihood

DRE digital rectal exam

EB empirical Bayes

ERSPC European Randomized Study of Screening for Prostate Cancer risk calculator 3

ESE estimated standard error

IVW inverse variance weighted estimator

MLE maximum likelihood estimation

MSE mean squared error

OCWE optimal covariates weighted estimator

PCPThg Prostate Cancer Prevention Trial (high-grade prostate cancer risk calculator)

PCA3 biomarker prostate cancer antigen 3

PSA prostate-specific antigen

SC-Learner selective coefficient learner

SD Monte Carlo standard deviation from 500 simulations

SSE sum of squared error

TRUS-PV transrectal ultrasound prostate volume

T2:ERG biomarker TMPRSS2:ERG

ABSTRACT

In the era of big data, it is becoming increasingly common for researchers to consider incorporating external information from large studies to improve the accuracy of statistical inference instead of relying on a modestly sized dataset collected internally. We consider a general statistical problem where there are some known regression models or risk calculators to predict an outcome of interest from a set of commonly used predictors. Different types of summary information are available for these external models. An internal modest-sized dataset containing individual-level data for the variables in the known models and some new variables is available for our current analysis. In all three chapters below, we consider different settings to achieve the same goal–to build an improved prediction model that includes the new variables, using both the internal individual-level data and summary information obtained from the known external model(s).

In Chapter 2, we focus on the simple case where there is only one large, well-characterized previous study from the external population. We propose a synthetic data approach, which first converts the external information into synthetic data, and then analyzes a combined dataset consisting of the observed internal data and the synthetic data. A theoretical justification and extensive simulation studies establish the efficiency gain and improved prediction performance of the proposed data integration method. We also illustrate that even under less restrictive requirements on the information that is available externally, the combined estimates have the same asymptotic properties as an alternative constraint maximum likelihood estimation approach.

In Chapter 3, we consider a more complicated but quite plausible situation where several external prediction models are available to aid inference and prediction for the internal study. We assume that each of the external studies developed a prediction model for the same outcome but may use a slightly different set of covariates. We propose a meta-inference framework using an empirical Bayes estimation approach, which adaptively combines the estimates from the external models. This adaptive approach diminishes the influence of information that is less compatible with the internal data while balancing the bias-variance trade-off. The estimators we proposed are more efficient than the naive analysis of the internal data.

In Chapter 4, we first extend the synthetic data method from Chapter 2 to accommodate the situation with multiple external prediction models, and further allow for heterogeneity of covariate effects across the external populations. Each external model could potentially be built on slightly

different subsets of covariates that are measured in the internal study. The proposed approach generates synthetic outcome data in each population, uses stacked multiple imputation to create a long dataset with complete covariate information, and finally analyzes the imputed data with weighted regression. Leveraging multiple sources of auxiliary information from a broad class of externally fitted predictive models or established risk calculators based on parametric regression or machine learning methods, this new strategy can make statistical inference more accurate for both the internal population and the external populations.

We evaluate the proposed methods through extensive simulations and apply them to improve models for predicting the risk of high-grade prostate cancer.

CHAPTER 1

Introduction

Increasingly, researchers are considering incorporating external information from large-scale studies to improve statistical inference rather than using the limited-sized data that are available to each investigator. Examples of this are borrowing strength from historical control data to leverage the treatment effect in small-sample clinical trials [Viele et al., 2014, Dejardin et al., 2018, Li and Song, 2020], combining separate probability samples [Bycroft, 2011, Yang and Kim, 2020] and incorporating external data sources for improved causal inference [Yang and Ding, 2020a]. However, challenges exist, such as data sharing, storage, and privacy issues to access publicly available individual-level large data, so often only the summary information is reported. Examples of such data sources include publications, online risk calculators, census data and population-based biobank data. Therefore, general frameworks that integrate the individual-level data and the summary-level external information are needed.

As a motivating example, it is common in predictive modeling that researchers want to include new predictors to update the traditional risk models in clinical biomedicine, such as adding genetic risk variants and mammographic density to the breast cancer risk calculator [Gail et al., 1989]. These traditional models are usually constructed from large datasets using principled statistical methods to predict a measure of risk or disease state, treating the patient characteristics as predictors. The patient characteristics, denoted by $\mathbf{X}$, can range from traditional epidemiologic, clinical and behavioral variables to well-known imaging, genetic and other molecular biomarkers. The predicted outcome variable Y, and the predictors $\mathbf{X}$, are often assumed to be connected through a regression model of the form $\mathrm{Y}|\mathbf{X}$. The individual-level original data that were used to construct this model are usually not available to the public, but what are accessible are certain forms of summary-level information. This information can be available in the form of coefficient estimates for the fitted model [Thompson et al., 2006, Roobol et al., 2012] or individual prediction probabilities, of which the underlying model may or may not be known, especially when the external information comes in the form of a "black box" algorithm, i.e. an algorithm that provides a predicted probability, but the underlying model is not necessarily simple or transparent or even known

[Stephan et al., 2003, Osareh and Shadgar, 2010, Estiri et al., 2021]. Furthermore, different external models predicting the same outcome may use different sets of predictors [Thompson et al., 2006, Roobol et al., 2012].

While these existing models are often based on traditional epidemiologic and behavioral risk factors and well-established biomarkers, wider availability of high throughput data and novel assay technologies are generating new candidate biomarkers, say B, for possible inclusion in existing risk prediction models. Due to the potential improvement of prediction accuracy of the current model, it is ideal to incorporate B into the well-established model $\mathrm{Y}|\mathbf{X}$, and construct an expanded prediction model of interest $\mathrm{Y}|\mathbf{X}, \mathrm{B}$. However, it is very likely that B and $\mathbf{X}$ are assessed only on participants in a study of moderate size and cannot be retrospectively measured on the much larger population used for $\mathrm{Y}|\mathbf{X}$ model. It is natural to consider using the information from the well-established model to increase the accuracy of the expanded model. This represents a general statistical challenge to build a good model for $\mathrm{Y}|\mathbf{X}, \mathrm{B}$ that uses both the known external information from the $\mathrm{Y}|\mathbf{X}$ model and the individual-level data from a small sample dataset of Y, $\mathbf{X}$ and B.

There exist proposals in the literature to incorporate external information into regression estimation. Imbens and Lancaster [1994] investigated how aggregate data (e.g. the population average of the response) could be used to improve maximum likelihood estimates in a regression model. Grill et al. [2015] proposed a simple method of incorporating new markers into an existing calculator via Bayes Theorem. Studies on this topic begin with incorporating external auxiliary information from large data, such as census or population-based biobank data, to improve the statistical inference of the internal study, assuming the model that relates the variables is fully or partially shared between data sources. The topic of whether distributions are similar between populations is called transportability [Bareinboim and Pearl, 2013] and is crucial to consider in the field of data integration. Qin [2000], Han and Lawless [2019] proposed possible solutions using empirical likelihood, Chatterjee et al. [2016], Cheng et al. [2018] demonstrated it from the perspective of constrained maximum likelihood while Cheng et al. [2019] proposed to use the Bayesian approach. The performance of various approaches was assessed in a simulation study by Grill et al. [2017]. Estes, Mukherjee, and Taylor [2017] then relaxed the transportability assumption of Chatterjee et al. [2016]'s by constructing an empirical Bayes estimator that protected against the potential bias.

Recent studies started to tackle challenges such as accommodating multiple external data sources and/or heterogeneity exists among data sources. Kundu, Tang, and Chatterjee [2019] extended the work of Chatterjee et al. [2016] to a meta-analysis setting. Chen et al. [2020] harmonized the difference of aggregate information among data sources through a penalty function while Yang and Ding [2020b] employed a sensitivity parameter to quantify such systematic differences. Moreover, there could be other sources of information variation across the models. For example, different external studies may use a different subset of covariates and the underlying prediction

model may be parametric or constructed by machine learning approaches. The summary-level information may contain estimated regression coefficients or fitted predictions.

In this dissertation, we consider a general statistical problem where there are some known regression models or risk calculators to predict an outcome of interest from a set of commonly used predictors. Different types of summary information are available for these external models. An internal modest-sized dataset containing individual-level data for the variables in the known models as well as some new variables is available for our current analysis. In all three chapters below, we consider different settings aiming to achieve the same goal–to build an improved prediction model that includes the new variables, using both the internal individual-level data and summary information obtained from the known external model(s).

In Chapter 2, we focus on the simple case where there is only one large, well-characterized previous study from the external population. We propose a synthetic data approach, which first converts the external information into synthetic data, and then analyzes a combined dataset consisting of the observed internal data and the synthetic data. A theoretical justification and extensive simulation studies establish the efficiency gain and improved prediction performance of the proposed data integration method. We also illustrate that even under less restrictive requirements on the information that is available externally, the combined estimates have the same asymptotic properties as an alternative constraint maximum likelihood estimation approach.

In Chapter 3, we consider a more complicated but quite plausible situation where several external prediction models are available to aid inference and prediction for the internal study. We assume that each of the external studies developed a prediction model for the same outcome but may use a slightly different set of covariates. We propose a meta-inference framework using an empirical Bayes estimation approach, which adaptively combines the estimates from the external models. This adaptive approach diminishes the influence of information that is less compatible with the internal data while balancing the bias-variance trade-off. The estimators we proposed are more efficient than the naive analysis of the internal data.

In Chapter 4, we first extend the synthetic data method from Chapter 2 to accommodate the situation with multiple external prediction models, and further allow for heterogeneity of covariate effects across the external populations. Each external model could potentially be built on slightly different subsets of covariates that are measured in the internal study. The proposed approach generates synthetic outcome data in each population, uses stacked multiple imputation to create a long dataset with complete covariate information, and finally analyzes the imputed data with weighted regression. Leveraging multiple sources of auxiliary information from a broad class of externally fitted predictive models or established risk calculators based on parametric regression or machine learning methods, this new strategy can make statistical inference more accurate for both the internal population and the external populations.

In all the chapters, we evaluate the proposed methods through extensive simulations and apply them to improve models for predicting the risk of high-grade prostate cancer.

CHAPTER 2

Synthetic Data Method to Incorporate External Information into a Current Study

2.1 Introduction

In Chapter 1, we introduced some of the existing methods to solve this genre of problem. In general, the constrained maximum likelihood (CML) approaches require a specific form for the external information, e.g. estimated coefficients from a correctly specified mean model and assumptions regarding the transportability of the distribution of $\mathrm{Y}, \mathbf{X}, \mathrm{B}$ across the internal and external sample. The CML approach proposed in Cheng et al. [2018] also requires the specification of a model for $\mathrm{B}|\mathbf{X}$ and relies on some parametric assumptions. Chatterjee et al. [2016] described a constrained semi-parametric maximum likelihood (CSPML) method by converting the external summary-level information into a constraint and then maximizing the internal data likelihood subject to this constraint. Although this CSPML approach does not require the $\mathrm{Y}|\mathbf{X}$ model to be correctly specified or a model for $\mathrm{B}|\mathbf{X}$, it requires the transportability of the joint distribution of $\mathrm{Y}, \mathrm{X}, \mathrm{B}$. Estes, Mukherjee, and Taylor [2017] and Cheng et al. [2018] have found that violation of this assumption and the small sample size in the internal data will cause sensitive and unstable estimation.

In this chapter, we propose a synthetic data framework as a more flexible alternative solution to the CML approach, motivated by methods developed in the survey methodology literature [Raghunathan et al., 2003, Reiter and Kinney, 2012, Reiter, 2002]. In this approach, synthetic data for $\mathbf{X}$ and Y are generated from the observed data and the $\mathrm{Y}|\mathbf{X}$ model, respectively, and then added to the observed data, then from this combined dataset a model for $\mathrm{Y}|\mathbf{X}, \mathrm{B}$ is built. Our method relaxes the requirement on the information that is available from the external model such that the only requirement is the ability to generate predictions of Y given $\mathbf{X}$, which can make use of external information that comes in the form of a "black box" algorithm, i.e., an algorithm that provides a predicted probability, but the underlying model is not necessarily simple or transparent or even known.

The following is the structure of the remainder of this chapter: in Section 2.2, we introduce the notation, assumptions and implementation of the proposed synthetic data method. In Section 2.3, under various simulation scenarios, we evaluate the performance of the synthetic data method. We demonstrate the proposed method through an application to the Prostate Cancer Prevention Trial data in Section 2.4. We provide some theoretical justification and insight for the synthetic data method in Section 2.5. In two special cases we show that with a very large number of synthetic observations, our approach gives identical asymptotic variances for the parameters of the $\mathrm{Y}|\mathbf{X}, \mathrm{B}$ model as the CML estimation approach that exists in the literature. Because the CML is a maximum likelihood estimator, it is optimal if the models are correctly specified. Since the synthetic data method has the same asymptotic variance, it can also be considered optimal. Concluding remarks are presented in Section 2.6.

2.2 Method

2.2.1 General Description of the Problem

Let Y denote the outcome of interest, which can be either continuous or binary. $\mathbf{X}$ is a set of p standard variables and let B denote a new biomarker. There are two populations, an external population for which we do not have individual-level data and an internal population for which we do have a dataset of size n with subject-level data. We will assume that the distributions of $\mathrm{Y}|\mathbf{X}, \mathrm{B}$ are the same in the two populations, and likewise for the $\mathrm{Y}|\mathbf{X}$ distribution. Our target of interest is the mean structure of $\mathrm{Y}|\mathbf{X}, \mathrm{B}$:

$$g[\mathrm{E}(\mathrm{Y}|\mathbf{X}, \mathrm{B})] = \gamma_0 + \gamma_{\mathrm{X}_1}\mathrm{X}_1 + ... + \gamma_{\mathrm{X}_\mathrm{p}}\mathrm{X}_\mathrm{p} + \gamma_\mathrm{B}\mathrm{B}, \tag{2.1}$$

where g is the known link function. We assume that a small dataset of size n with variables $\mathrm{Y}, \mathbf{X}$ and a new covariate B is available to us for building the model of interest.

We assume a large, well-characterized previous study from the external population describes the provided information on the calculated distribution of $\mathrm{Y}|\mathbf{X}$. This information can come in various forms, including partial or full knowledge of the distribution of the $\mathrm{Y}|\mathbf{X}$ model.

2.2.2 Synthetic Data Method

We propose an algorithmic approach that can produce synthetic data on $(\mathrm{Y}, \mathbf{X}, \mathrm{B})$, by using the combination of the available information from the established model and the observations from the current dataset. The synthetic data would incorporate the external information as well as enlarge

the sample size, and thus it helps improve the inference about coefficients γ in model 2.1, compared to just analyzing the small dataset based on the observed data.

The synthetic data approach consists of creating m additional synthetic data observations, and then analyzing the combined dataset of size n+m to estimate the parameters of model 2.1. The synthetic data are created in two steps as shown in Figure 2.1. In step 1, we replicate $\mathbf{X}$ a large number (say S) times in blocks of n rows to create m = nS additional records. In step 2, we generate pseudo data called Y^* from the known $Y|\mathbf{X}$ distribution for these new m records. Finally, we combine the synthetic observations with the original dataset, and we note that the combined data will now have missing values of B for m observations. The combined data is then analyzed to give an estimate of γ.

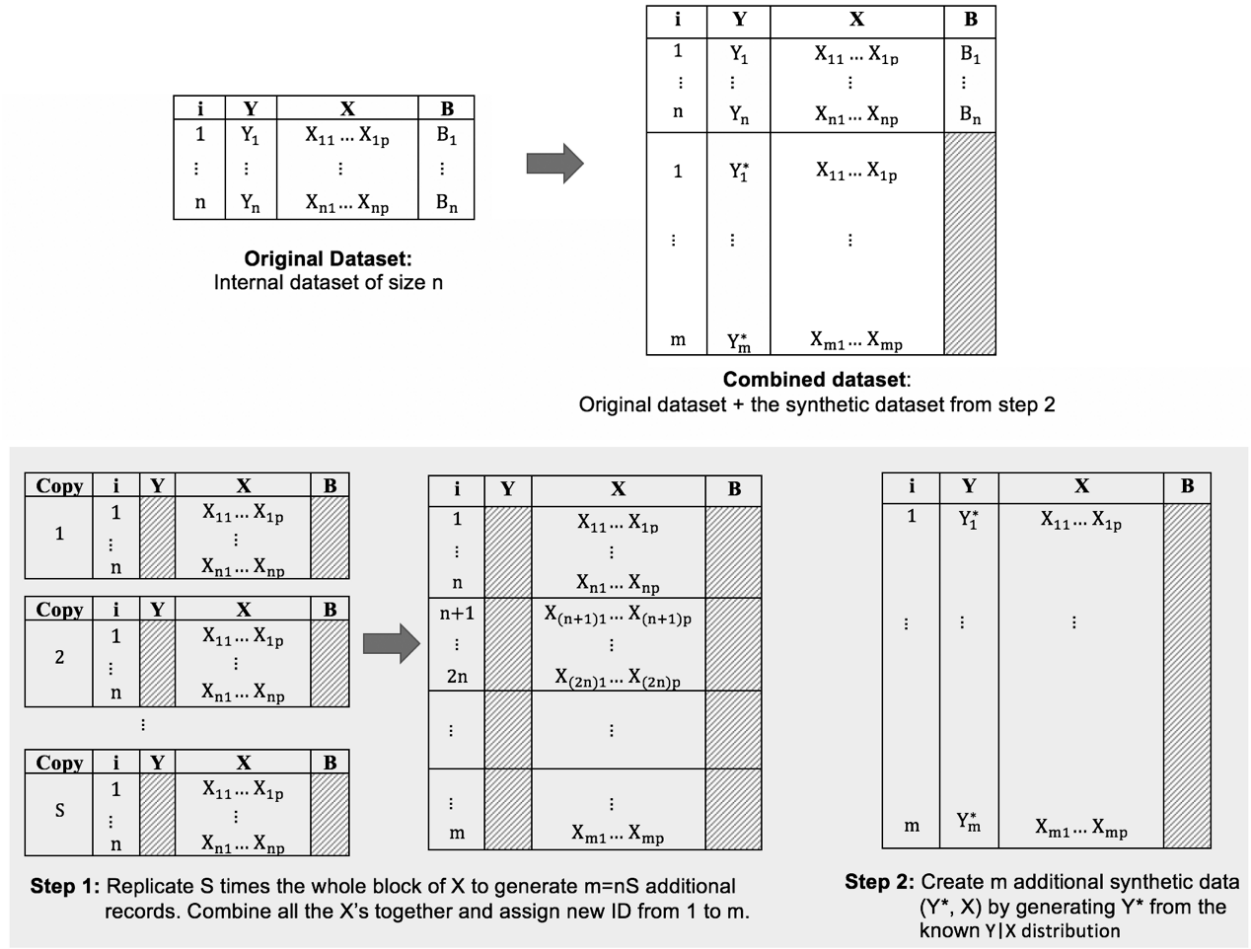


Figure 2.1: Two steps to create the synthetic data

There may be different ways in which the combined data can be analyzed. In Section 5 we present two special cases for which a closed-form MLE of γ exists for the combined dataset of size n+m. In other cases, like the simulation study settings in Section 3, no closed-form solution

for the MLE of $\boldsymbol{\gamma}$ exists, and our proposed approach to deal with missing data is to use multiple imputation to impute the m missing values of B. Multiple imputation is a general procedure for analyzing datasets with missing values. It consists of defining a procedure to fill in the missing values, then applying that procedure many times to create many separate complete datasets. Each completed dataset is then analyzed and the results of these separate analyses are combined to give final estimates. In this particular case the multiple imputation approach requires us to specify a parametric model $(\mathrm{B}|\mathbf{X}, \mathrm{Y})$, from which we draw 50 values of B to give 50 completed datasets. Then we fit model 2.1 for each complete data (Y, $\mathbf{X}$, B) of size n+m. We then average the estimates of $\boldsymbol{\gamma}$ from the 50 complete datasets, and compute the total variance using Rubin's rules [Rubin, 1987]. We then proceed with inference.

Multiple imputation has the additional advantage of being able to handle multiple biomarkers in B, some of which may be discrete and some continuous. It also allows for flexible structure for the conditional mean model for each biomarker in B given all other variables in the dataset, such as the possibility to incorporate non-linearity and interactions. For implementing multiple imputation, we use the R package MICE [Van Buuren and Oudshoom, 2011]. We use the function $mice$ with imputation algorithm *logreg* (the Bayesian logistic regression model with flat prior) for the imputation of a binary B and the imputation algorithm $norm$ (the Bayesian linear regression model) for the imputation of a continuous B. In the situation in which there are multiple B's, say B_1 and B_2, imputations are done sequentially. That is, first draw B_1 from the $\mathrm{B}_1|\mathbf{X}, \mathrm{Y}, \mathrm{B}_2$ distribution, then draw B_2 from the $\mathrm{B}_2|\mathbf{X}, \mathrm{Y}, \mathrm{B}_1$ distribution, and iterate between B_1 and B_2.

2.3 Simulation Study

To assess the performance of the proposed synthetic data method for both estimation and prediction, we conduct simulation studies under four different scenarios. Each scenario has a different true distribution for $\mathrm{Y}|\mathbf{X}, \mathrm{B}$ and for $\mathrm{B}|\mathbf{X}$ for the internal data. For both the outcome and the predictors, we consider both continuous and binary variables to illustrate the computational implementation in a range of situations. We also consider the situation of multiple B's to evaluate the applicability of the synthetic method in multi-dimensional cases. In some cases a misspecified imputation model is used within the synthetic data approach, thus allowing us to evaluate the robustness of the method. Only in special cases (see Section 2.5) can we provide a theoretical justification for the synthetic data approach, thus the simulations are intended to provide numerical properties of the synthetic data approach in situations where the relevant theoretical properties are not yet available.

In real situations, we expect a moderate number of X variables, and their joint distribution could be quite complex with skew distributions and correlations between different X's. To achieve this

we adopted a procedure of generating X's as described in Xu, Daniels, and Winterstein [2016]. We generate 9 correlated X's in each of the four simulation scenarios as described below:

$u_j \sim$ N(0, 1), $\quad j$=1...5, $\quad \mathrm{X}_1 \sim$ N(0, 1)

$$\mathrm{X}_j|u_1,...,u_5 = \begin{cases} u_1 u_j + \epsilon_j, \epsilon_j \sim \frac{2}{3}\mathrm{N}_+(0,0.2) + \frac{1}{3}\mathrm{N}_-(0,0.1) & \mathrm{j} = 2,...,5 \\ u_2 u_{j-3} + \epsilon_j, \epsilon_j \sim \frac{1}{4}\mathrm{N}_+(0,0.1) + \frac{3}{4}\mathrm{N}_-(0,0.3) & \mathrm{j} = 6,...,8 \\ u_3 u_{j-5} + \epsilon_j, \epsilon_j \sim \frac{4}{5}\mathrm{N}_+(0,0.4) + \frac{1}{5}\mathrm{N}_-(0,0.1) & \mathrm{j} = 9 \end{cases}$$

where N_+ and N_- represent half normal distributions, and either one or the other is selected with the shown probability. We then generate B from the $\mathrm{B}|\mathbf{X}$ distribution, and finally generate Y from the $\mathrm{Y}|\mathbf{X}, \mathrm{B}$ distribution. For scenario 1 (where Y is continuous, as described below) the form of the external model for $\mathrm{Y}|\mathbf{X}$ is readily available. For cases (scenarios 2, 3, and 4 where Y is binary, as described below) where the closed-form of model $\mathrm{Y}|\mathbf{X}$ is not available, we numerically derive the external model $\mathrm{Y}|\mathbf{X}$. Specifically, we generate an independent dataset of $(\mathrm{Y}, \mathbf{X}, \mathrm{B})$ of size 10000 and fit a linear or logistic regression model g[E($\mathrm{Y}|\mathbf{X}$)] depending on the type of Y. The estimated coefficients of this model serve as the external information we obtained from the established model $\mathrm{Y}|\mathbf{X}$.

For each simulation scenario, we first simulate 500 datasets of size n. Then we create the synthetic data following the steps introduced in Section 2, and combine them with the original data to get 500 datasets of size n+m with m missing B values. For each simulated dataset, we create 50 complete datasets by imputing the missing B values given Y and $\mathbf{X}$. In all four scenarios, we use linear additive models for imputing from the $\mathrm{B}|\mathbf{X}, \mathrm{Y}$ distribution, without including any interaction terms. We compare the results of the synthetic data method to the direct MLE, which uses the complete dataset of size n, in terms of estimation accuracy and prediction ability. We report the average estimated coefficients, standard deviation and 95% coverage rate for $\hat{\boldsymbol{\gamma}}$. To measure the predictive performance, we generate a new dataset of size 1500 for each scenario, and evaluate the prediction $\hat{\mathrm{Y}}_\mathrm{i}$ on this new dataset. In the new validation dataset, let $\hat{\mathrm{p}}$ or $\bar{\mathrm{Y}}$ denote the average of the generated Y values. For the continuous Y, we use the mean squared error (MSE) defined as $\sum_{\mathrm{i}=1}^{1500}(\hat{\mathrm{Y}}_\mathrm{i} - \bar{\mathrm{Y}})^2 / \sum_{\mathrm{i}=1}^{1500}(\mathrm{Y}_\mathrm{i} - \bar{\mathrm{Y}})^2$. For binary Y, we use AUC and scaled Brier score, defined as $\sum_{\mathrm{i}=1}^{1500}(\mathrm{Y}_\mathrm{i} - \hat{\mathrm{Y}}_\mathrm{i})^2 / \sum_{\mathrm{i}=1}^{1500}(\mathrm{Y}_\mathrm{i} - \hat{\mathrm{p}})^2$, as measures of discrimination and calibration.

The four simulation scenarios and results are described as follows:

- **Scenario 1:** Y and B are Gaussian distributed. The true model of $\mathrm{Y}|\mathbf{X}, \mathrm{B}$ is $\mathrm{Y}_\mathrm{i}|\mathbf{X_i}, \mathrm{B}_\mathrm{i}$ = $0.5\sum_{\mathrm{j}=1}^{9} \mathrm{X}_{\mathrm{ji}} + \mathrm{B}_\mathrm{i}$ + e_i, $e_i \sim$ N(0, 3), and B_i is simulated as $\mathrm{B}_\mathrm{i} = 0.2\sum_{\mathrm{j}=1}^{9} \mathrm{X}_{\mathrm{ji}}$ + f_i, $f_i \sim$ N(0, 1). The reduced $\mathrm{Y}|\mathbf{X}$ model is $\mathrm{Y}_\mathrm{i} = 0.7\sum_{\mathrm{j}=1}^{9} \mathrm{X}_{\mathrm{ji}} + \mathrm{B}_\mathrm{i}$ + k_i, $k_i \sim$ N(0, 4). The current data sample size n = 200, replication number S = 10, and thus the synthetic data sample size m = nS = 2000.

Results: The results in Table 2.1 of scenario 1 show that compared to the direct MLE, the synthetic data method leads to an obvious reduction in the standard deviation of γ_{X}'s and good coverage rates of confidence intervals. In addition, it can reduce the MSE closer to the true value by 50%.

Table 2.1: Simulation results for scenario 1 with Gaussian Y, one Gaussian B and nine correlated X's. For each method, we report mean (SD) [95 % coverage rate] and MSE across 500 simulated datasets.

	Not including B	True value	Direct MLE	Synthetic Data Method
γ_0	0	0	-0.04 (0.14) [97%]	0.00 (0.18) [96%]
γ_{X_1}	0.7	0.5	0.50 (0.13) [94%]	0.51 (0.08) [95%]
γ_{X_2}	0.7	0.5	0.48 (0.13) [94%]	0.50 (0.07) [95%]
γ_{X_3}	0.7	0.5	0.49 (0.12) [95%]	0.50 (0.07) [95%]
γ_{X_4}	0.7	0.5	0.50 (0.11) [98%]	0.50 (0.07) [95%]
γ_{X_5}	0.7	0.5	0.50 (0.13) [93%]	0.50 (0.08) [93%]
γ_{X_6}	0.7	0.5	0.50 (0.12) [94%]	0.50 (0.07) [95%]
γ_{X_7}	0.7	0.5	0.50 (0.12) [94%]	0.50 (0.07) [95%]
γ_{X_8}	0.7	0.5	0.50 (0.11) [97%]	0.50 (0.07) [95%]
γ_{X_9}	0.7	0.5	0.50 (0.13) [94%]	0.51 (0.07) [95%]
γ_{B}	-	0.5	1.00 (0.13) [95%]	1.00 (0.11) [95%]
MSE	0.464	0.334	0.355	0.345

- **Scenario 2:** Y is binary and B is Gaussian distributed. The true model of $\mathrm{Y}|\mathbf{X}, \mathrm{B}$ is $logit[\Pr(\mathrm{Y_i} = 1|\mathbf{X_i}, \mathrm{B_i})] = 1+2\sum_{\mathrm{j}=1}^{9} \mathrm{X_{ji}} - 3\mathrm{B_i}$, which gives the prevalence $\Pr(\mathrm{Y} = 1) \approx 0.67$. $\mathrm{B_i}$ is simulated as $\mathrm{B_i} = 0.5\sum_{\mathrm{j}=1}^{9} \mathrm{X_{ji}} + e_i$, $e_i \sim$ N(0, 0.1). The current data sample size n = 400, S = 10, and m = nS = 4000.

  Results: As shown in Table 2.2, including B into the regression model can reduce the scaled Brier score by 15%, and improve the AUC by 9%.

Table 2.2: Simulation results for scenario 2 with binary Y, one Gaussian B and nine correlated X's. For each method, we report mean (SD) [95 $\%$ coverage rate], average scaled Brier score and AUC across 500 simulated datasets

	Not including B	True value	Direct MLE	Synthetic Data Method
γ_0	0.849	1	1.00 (0.17) [94%]	0.96 (0.08) [92%]
γ_{X_1}	0.435	2	2.01 (0.29) [96%]	1.92 (0.24) [94%]
γ_{X_2}	0.432	2	2.00 (0.28) [96%]	1.90 (0.23) [94%]
γ_{X_3}	0.437	2	2.01 (0.28) [95%]	1.90 (0.23) [95%]
γ_{X_4}	0.433	2	2.01 (0.30) [96%]	1.91 (0.24) [95%]
γ_{X_5}	0.422	2	2.02 (0.28) [95%]	1.90 (0.24) [93%]
γ_{X_6}	0.421	2	2.01 (0.28) [97%]	1.89 (0.23) [93%]
γ_{X_7}	0.431	2	2.01 (0.29) [96%]	1.91 (0.23) [94%]
γ_{X_8}	0.415	2	2.00 (0.27) [97%]	1.89 (0.23) [94%]
γ_{X_9}	0.445	2	2.01 (0.29) [96%]	1.92 (0.23) [95%]
γ_{B}	-	-3	-3.02 (0.45) [97%]	-2.85 (0.43) [95%]
Scaled Brier score	0.801	0.680	0.702	0.686
AUC	0.767	0.837	0.828	0.835

- **Scenario 3:** Y and B are both binary. The true model of $\mathrm{Y}|\mathbf{X}, \mathrm{B}$ is $logit[\Pr(\mathrm{Y_i} = 1|\mathbf{X_i}, \mathrm{B_i})] = -1 + 0.2\sum_{\mathrm{j}=1}^{4}\mathrm{X_{ji}} - 0.2\sum_{\mathrm{j}=5}^{7}\mathrm{X_{ji}} - 0.5\sum_{\mathrm{j}=8}^{9}\mathrm{X_{ji}} + 1.5\mathrm{B_i}$, and $\mathrm{B_i}$ is simulated as $logit[\Pr(\mathrm{B_i} = 1|\mathbf{X_i})] = -0.5 + 0.5\sum_{\mathrm{j}=1}^{5}\mathrm{X_{ji}} - \sum_{\mathrm{j}=6}^{9}\mathrm{X_{ji}}$. The prevalence of Y and B are around 0.5 and 0.55, respectively, where n = 400, S = 8, and m = nS = 3200.

  Results: The simulation results in Table 2.3 for scenario 3, where both Y and B are binary, show that including B in the regression model can reduce the scaled Brier score by 10.8%, and increase the AUC by 5%.

Table 2.3: Simulation results for scenario 3 with binary Y, one binary B and nine correlated X's. For each method, we report mean (SD) [95 $\%$ coverage rate], average scaled Brier score and AUC across 500 simulated datasets

	Not including B	True value	Direct MLE	Synthetic Data Method
γ_0	-0.328	-1	-1.00 (0.21) [94%]	-1.00 (0.15) [96%]
γ_{X_1}	0.305	0.2	0.20 (0.13) [94%]	0.20 (0.05) [96%]
γ_{X_2}	0.318	0.2	0.21 (0.13) [95%]	0.21 (0.05) [94%]
γ_{X_3}	0.296	0.2	0.20 (0.13) [95%]	0.21 (0.05) [95%]
γ_{X_4}	0.296	0.2	0.19 (0.13) [93%]	0.20 (0.05) [95%]
γ_{X_5}	-0.066	-0.2	-0.21 (0.13) [94%]	-0.19 (0.05) [96%]
γ_{X_6}	-0.405	-0.2	-0.20 (0.13) [96%]	-0.20 (0.06) [96%]
γ_{X_7}	-0.420	-0.2	-0.19 (0.13) [96%]	-0.21 (0.06) [95%]
γ_{X_8}	-0.698	-0.5	-0.50 (0.15) [93%]	-0.51 (0.06) [96%]
γ_{X_9}	-0.713	-0.5	-0.51 (0.14) [95%]	-0.52 (0.06) [94%]
γ_{B}	-	1.5	1.50 (0.28) [96%]	1.49 (0.28) [95%]
Scaled Brier score	0.750	0.666	0.687	0.669
AUC	0.789	0.833	0.823	0.831

- **Scenario 4:** Y is binary and two mixed types of B are included, one binary and another Gaussian. The true model of $\mathrm{Y}|\mathbf{X}, \mathrm{B}_1, \mathrm{B}_2$ is $logit[\Pr(\mathrm{Y_i} = 1|\mathbf{X_i}, \mathrm{B_{1i}}, \mathrm{B_{2i}})] = -0.2 - 0.2\mathrm{X}_1 + 0.2\sum_{\mathrm{j=2}}^{3}\mathrm{X_{ji}} + 0.1\sum_{\mathrm{j=4}}^{5}\mathrm{X_{ji}} - 0.1\mathrm{X}_6 - 0.3\mathrm{X}_7 + 0.3\sum_{\mathrm{j=8}}^{9}\mathrm{X_{ji}} + 2\mathrm{B_{1i}} - \mathrm{B_{2i}}$, of which the prevalence P(Y=1) is approximately 0.53. The binary $\mathrm{B_{1i}}$ is simulated as $logit[\Pr(\mathrm{B_{1i}} = 1|\mathbf{X_i})] = -0.5 + 0.5\sum_{\mathrm{j=1}}^{5}\mathrm{X_{ji}} - \sum_{\mathrm{j=6}}^{9}\mathrm{X_{ji}}$, which gives the prevalence $\Pr(\mathrm{B}_1 = 1) \approx 0.56$. The Gaussian $\mathrm{B_{2i}}$ is simulated as $\mathrm{B_{2i}} = 0.3\sum_{\mathrm{j=1}}^{2}\mathrm{X_{ji}} - 0.2\sum_{\mathrm{j=3}}^{4}\mathrm{X_{ji}} + 0.5\sum_{\mathrm{j=5}}^{7}\mathrm{X_{ji}} - 0.5\sum_{\mathrm{j=8}}^{9}\mathrm{X_{ji}} + e_i$, $e_i \sim \mathrm{N}(0, 0.1)$. In this scenario, n = 400, S = 8, and m = nS = 3200.

  Results: The results in Table 2.4 show that the synthetic data method improves the scaled Brier score and AUC compared to the MLE, where these metrics almost attain the best possible values listed in the true value column. In addition, the coverage rates of the confidence intervals for the γ's close to 95% as desired.

Table 2.4: Simulation results for scenario 4 with binary Y, binary $\mathrm{B1}$ and continuous $\mathrm{B2}$ and nine correlated X's. For each method, we report mean (SD) [95 $\%$ coverage rate], average scaled Brier score and AUC across 500 simulated datasets

	Not including B	True value	Direct MLE	Synthetic Data Method
γ_0	0.250	-0.2	-0.197 (0.21) [96%]	-0.13 (0.15) [93%]
γ_{X_1}	0.441	-0.2	-0.19 (0.19) [96%]	-0.14 (0.13) [94%]
γ_{X_2}	0.779	0.2	0.21 (0.21) [94%]	0.23 (0.13) [94%]
γ_{X_3}	-0.132	0.2	0.19 (0.17) [95%]	0.17 (0.10) [94%]
γ_{X_4}	-0.218	0.1	0.09 (0.17) [95%]	0.08 (0.10) [95%]
γ_{X_5}	1.047	0.1	0.10 (0.26) [96%]	0.15 (0.21) [95%]
γ_{X_6}	0.705	-0.1	-0.10 (0.28) [94%]	-0.06 (0.21) [94%]
γ_{X_7}	0.529	-0.3	-0.29 (0.27) [94%]	-0.25 (0.21) [93%]
γ_{X_8}	-0.750	0.3	0.29 (0.25) [97%]	0.23 (0.21) [95%]
γ_{X_9}	-0.757	0.3	0.30 (0.25) [96%]	0.22 (0.21) [94%]
γ_{B1}	-	1	0.996 (0.31) [96%]	0.88 (0.30) [93%]
γ_{B2}	-	2	1.99 (0.45) [95%]	1.83 (0.43) [93%]
Scaled Brier score	0.637	0.575	0.598	0.582
AUC	0.849	0.876	0.868	0.873

In summary, the results of simulation studies show that: (1) the synthetic data method can improve the efficiency of estimating γ_{X}'s while reduce the MSE of the predictions and increase the AUC for binary Y; (2) In scenario 1 where the imputation model $\mathrm{B}|\mathbf{X}, \mathrm{Y}$ is correctly specified, the estimates of γ_{X}'s and γ_{B} are unbiased; (3) In scenarios 2, 3 and 4 where the imputation model $\mathrm{B}|\mathbf{X}, \mathrm{Y}$ is misspecified, despite the improved predictive performance, there is some bias when estimating γ_{X}'s and γ_{B}. In future work, we will investigate if we can achieve even further improvements in performance using alternative or more flexible or more nonparametric approaches for imputing B.

2.4 Prostate Cancer Prevention Trial Data Example

To assess the performance of the synthetic data method in a real data example, we apply it to the Prostate Cancer Prevention Trial calculator. The high-grade prostate cancer calculator (PCPThg) [Thompson et al., 2006], predicts the probability of high-grade prostate cancer derived from a logistic regression based on standard clinical variables – PSA level, age, DRE findings, prior biopsy result and ethnicity. The equation for the model is:

$$logit(\mathrm{p_i}) = -6.25 + 0.03\mathrm{age_i} + 0.96\mathrm{race_i} + 1.29\log(\mathrm{PSA_i}) + 1.00\mathrm{DRE_i} - 0.36\mathrm{biopsy_i}. \quad (2.2)$$

where p_i is the probability of observing high-grade prostate cancer for subject i given covariates. A detailed description of the calculator and the external and internal and a validation dataset are given in Tomlins et al. [2015] and Cheng et al. [2018]. We consider incorporating two additional biomarkers that have been shown to be predictive of prostate cancer into model 2.2. One is prostate cancer antigen 3 (PCA3), a continuous variable, and the other is the indicator variable of TMPRSS2:ERG (T2:ERG) gene fusions. We consider 3 different expanded models, one with the addition of PCA3 only, one with the addition of T2:ERG only and one with the addition of both PCA3 and T2:ERG. To compare the coefficient estimation across methods, we show the estimated coefficients and standard errors in Table 2.5 from 679 observations in the internal dataset. To compare prediction power, we calculate the scaled Brier Score and the AUC based on an independent validation dataset with 1,218 observations.

Table 2.5: Results of the expanded PCPThg model. For each method, point estimate (standard error) from the internal dataset, and the scaled Brier score and the AUC from the validation dataset. The sample size of the internal dataset is 679. The sample size of the validation dataset is 1,218. Replicate number S=10 giving m=6,790 in the synthetic data method.

Model	PSA	Age	DRE findings	Prior biopsy history	Race	PCA3	T2:ERG	Scaled Brier Score	AUC
Original PCPThg	1.29	0.031	1.00	-0.36	0.96	–	–	0.933	0.707
Estimated PCPThg	1.06 (0.18)	0.033 (0.012)	1.15 (0.26)	-1.44 (0.27)	0.44 (0.29)	–	–	0.975	0.716
Expanded model with PCA3 score									
Direct regression*	0.97 (0.19)	0.009 (0.013)	1.06 (0.27)	-1.27 (0.27)	0.05 (0.31)	0.56 (0.08)	–	0.953	0.767
Synthetic data method	1.30 (0.08)	0.012 (0.006)	0.91 (0.13)	-0.56 (0.12)	0.50 (0.14)	0.57 (0.08)	–	0.878	0.765
CSPML	1.22 (0.08)	0.007 (0.005)	0.86 (0.10)	-0.20 (0.08)	0.58 (0.11)	0.56 (0.097)	–	0.888	0.759
Expanded model with binary T2:ERG									
Direct regression*	0.98 (0.18)	0.032 (0.012)	1.02 (0.26)	-1.41 (0.27)	0.57 (0.29)	–	0.76 (0.20)	0.930	0.744
Synthetic data method	1.21 (0.07)	0.030 (0.005)	0.96 (0.10)	-0.59 (0.09)	0.99 (0.11)	–	0.76 (0.22)	0.897	0.741
CSPML	1.14 (0.07)	0.032 (0.004)	1.06 (0.14)	-0.52 (0.11)	0.80 (0.17)	–	0.72 (0.20)	0.931	0.742
Expanded model with PCA3 score and binary T2:ERG									
Direct regression*	0.94 (0.19)	0.010 (0.010)	1.00 (0.28)	-1.27 (0.28)	0.15 (0.31)	0.52 (0.08)	0.47 (0.21)	0.928	0.776
Synthetic data method	1.23 (0.09)	0.008 (0.007)	0.83 (0.13)	-0.53 (0.11)	0.63 (0.15)	0.55 (0.10)	0.45 (0.20)	0.867	0.773
CSPML	1.20 (0.08)	0.008 (0.005)	0.78 (0.11)	-0.21 (0.09)	0.67 (0.12)	0.52 (0.10)	0.48 (0.27)	0.879	0.769

*Firth corrected MLE is used

As shown in Table 2.5, for both of the expanded PCPThg models incorporating PCA3 score or binary T2:ERG, if we compare the standard errors across different methods, it is easily seen that the synthetic data method can reduce the standard errors of regression coefficients compared to direct regression by at least 50%.

The expanded PCPThg model incorporating both PCA3 score and binary T2:ERG fitted to the training dataset again shows that the proposed method can reduce the standard errors of regression coefficients compared to direct regression. The results in Table 2.5 show no improvement in AUC from using the synthetic data approach compared to direct MLE, but noticeable improvement in

the scaled Brier score.

We also include in Table 2.5 the estimates from applying the CSPML, a published method that can be applied in this case. We see it gives similar predictive performance as the synthetic data method, but the estimated coefficients differ.

2.5 Algebraic Justification in Two Special Cases

2.5.1 Estimation and Variance of γ

To establish that the synthetic data approach is asymptotically as efficient as CML approaches, we consider two special cases where closed-form results of MLE for the combined dataset of size n+m in the synthetic data approach are available, so multiple imputation does not need to be used. For these cases, we compare the synthetic data approach to the basic CML method [Cheng et al., 2018] and the CSPML method [Chatterjee et al., 2016]. These two maximum likelihood approaches are optimal based on their assumptions. The standard maximal likelihood approach based on just the observed data without incorporating external information is also provided for reference comparison. For each approach, we derive the explicit formulas for the asymptotic variance of estimated coefficients, namely, $\hat{\gamma}$ in model (2.1).

The rationale for studying these two examples in depth is to establish some theoretical underpinning for the synthetic data approach. Given its broad applicability to other more general situations with a mixed set of continuous and categorical multivariable predictors in $\mathbf{X}$ and B, a justification in simpler cases that can be studied analytically makes the approach more plausible in other situations where studying the analytical properties is not immediate.

As needed, we will be considering three different likelihoods, one based on the distribution $\mathrm{Y}|\mathbf{X}, \mathrm{B}$, one based on the distribution $(\mathrm{Y}, \mathrm{B})|\mathbf{X}$ and one based on the joint distribution of $\mathrm{Y}, \mathbf{X}$ and B. When writing distributions, we will include the parameters when necessary, e.g. $f(\mathrm{Y}|\mathbf{X}, \mathrm{B}; \boldsymbol{\gamma})$, but parameters will be excluded when not necessary.

For this study we either know the full form of the distribution of $\mathrm{Y}|\mathbf{X}$, the mean of which may be characterized by a linear combination of $\mathbf{X}$'s, as given in equation 2.3, with known $\boldsymbol{\beta}$'s and a known link function g_1:

$$g_1[\mathrm{E}(\mathrm{Y}|\mathbf{X})] = \beta_0 + \beta_1 \mathrm{X}_1 + ... + \beta_\mathrm{p} \mathrm{X}_\mathrm{p}. \tag{2.3}$$

or we just know the mean structure but not the full distribution of $\mathrm{Y}|\mathbf{X}$.

As mentioned earlier, our interest lies in building the mean structure of the $\mathrm{Y}|\mathbf{X}, \mathrm{B}$ distribution as given in model 2.1. For some approaches, we will also need to consider the relationship between

$\mathbf{X}$ and B for which we specify a model, the mean of which is given by

$$g_2[\mathrm{E}(\mathrm{B}|\mathbf{X})] = \theta_0 + \theta_1 \mathrm{X}_1 + ... + \theta_\mathrm{p} \mathrm{X}_\mathrm{p}. \tag{2.4}$$

We note that for all of these models there may be additional parameters necessary to define the full distributions (e.g. the variance σ^2_β for Gaussian Y). But for ease of notation we will not include these additional parameters unless it is necessary, thus we denote the distributions as $f(\mathrm{Y}|\mathbf{X};\boldsymbol{\beta})$, $f(\mathrm{Y}|\mathbf{X},\mathrm{B};\boldsymbol{\gamma})$ and $f(\mathrm{B}|\mathbf{X};\boldsymbol{\theta})$.

As a reference comparison, we will also present results for standard MLE on a complete dataset of size n. In this approach, we estimate the parameters of model 2.1 using the internal dataset of $\mathrm{Y}, \mathbf{X}$, and B without taking the external summary-level information into account. We obtain the estimates by maximizing the likelihood $\prod_{\mathrm{i}=1}^{\mathrm{n}} f(\mathrm{Y_i}|\mathbf{X_i}, \mathrm{B_i};\boldsymbol{\gamma})$ over $\boldsymbol{\gamma}$. Then the asymptotic covariance matrix of $\hat{\boldsymbol{\gamma}}$ is obtained from the inverse of the Fisher information matrix.

- *Approach 1: The synthetic data method.* In special cases in which a direct solution is possible, the likelihood for a dataset of size n+m is $\prod_{\mathrm{i}=1}^{\mathrm{n}} f(\mathrm{Y_i}, \mathrm{B_i}|\mathbf{X_i}) \prod_{\mathrm{i}=\mathrm{n}+1}^{\mathrm{n+m}} f(\mathrm{Y_i}|\mathbf{X_i})$, and can also be written as $\prod_{\mathrm{i}=1}^{\mathrm{n}} f(\mathrm{B_i}|\mathrm{Y_i}, \mathbf{X_i}) \prod_{\mathrm{i}=1}^{\mathrm{n+m}} f(\mathrm{Y_i}|\mathbf{X_i})$. This likelihood is then maximized over $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ to obtain the MLE and the asymptotic variance is obtained from the inverse of the Fisher information.

- *Approach 2: CML on a complete dataset of size n.* For this approach we posit a model $f(\mathrm{B}|\mathbf{X};\boldsymbol{\theta})$ then maximize the likelihood $\prod_{\mathrm{i}=1}^{\mathrm{n}} f(\mathrm{Y_i}, \mathrm{B_i}|\mathbf{X_i})$, which can be written as

  $$\prod_{i=1}^{n} f(\mathrm{Y_i}|\mathbf{X_i}, \mathrm{B_i};\boldsymbol{\gamma}) f(\mathrm{B_i}|\mathbf{X_i};\boldsymbol{\theta})$$

  subject to a constraint on the parameters that is derived from the external information. The equality $f(\mathrm{Y}|\mathbf{X};\boldsymbol{\beta}) = \int f(\mathrm{Y}|\mathbf{X},\mathrm{B};\boldsymbol{\gamma}) f(\mathrm{B}|\mathbf{X};\boldsymbol{\theta}) d\mathrm{B}$ gives a relationship between the unknown parameters $\boldsymbol{\gamma}$, $\boldsymbol{\theta}$ and the known parameter $\boldsymbol{\beta}$. Assuming $\boldsymbol{\theta}$ can be written as a function of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, i.e. as $\boldsymbol{\theta}(\boldsymbol{\gamma}, \boldsymbol{\beta})$, then since $\boldsymbol{\beta}$ is known the optimization problem becomes an unconstrained optimization problem, specifically maximization of

  $$\prod_{\mathrm{i}=1}^{\mathrm{n}} f(\mathrm{Y_i}|\mathbf{X_i}, \mathrm{B_i};\boldsymbol{\gamma}) f(\mathrm{B_i}|\mathbf{X_i};\boldsymbol{\theta}(\boldsymbol{\gamma}, \boldsymbol{\beta} = \boldsymbol{\beta}^*))$$

  with respect to $\boldsymbol{\gamma}$ using the known value $\boldsymbol{\beta}^*$ of $\boldsymbol{\beta}$. We consider two variations of the CML method: Approach 2.1 where only the coefficients $\boldsymbol{\beta}$ are known, and Approach 2.2 where

both $\boldsymbol{\beta}$ and its variance σ_β^2 are known.

- *Approach 3: CSPML method applied to a dataset of size n.* For this method, the estimates are obtained by maximizing the likelihood $\prod_{\mathrm{i}=1}^{\mathrm{n}} f(\mathrm{Y_i}, \mathbf{X_i}, \mathrm{B_i})$ over $\boldsymbol{\gamma}$ and the empirical distribution of ($\mathbf{X}$, B), subject to a constraint [Chatterjee et al., 2016]. In this approach the distribution of ($\mathbf{X}$, B) is treated nonparametrically, and the constraint is derived from the integrated score equation of model 2.3. In this case the constraint is $\mathrm{E_{XB}}\big[\mathrm{E_{Y|XB}}[\frac{\partial}{\partial \boldsymbol{\beta}} log\{f(\mathrm{Y}|\mathbf{X};\boldsymbol{\beta})\}]\big] = 0$. The constrained optimization problem is implemented via Lagrange multipliers and gives both an estimate of $\boldsymbol{\gamma}$ and the non-parametric MLE of the distribution of ($\mathbf{X}$, B). The asymptotic variance of $\hat{\boldsymbol{\gamma}}$ for this approach is given by $(\mathbf{I} + \mathbf{C}\mathbf{L}^{-1}\mathbf{C}^{\mathrm{T}})^{-1}$, where

$$\begin{aligned}
\mathbf{I} &= \mathrm{E_{XB}}\big[\mathrm{E_{Y|XB}}[-\frac{\partial^2}{\partial \boldsymbol{\gamma}^2} log\{f(\mathrm{Y}|\mathbf{X},\mathrm{B};\boldsymbol{\gamma})\}]\big] \\
\mathbf{C} &= \mathrm{E_{XB}}\big[\mathrm{E_{Y|XB}}[\frac{\partial}{\partial \boldsymbol{\gamma}} log\{f(\mathrm{Y}|\mathbf{X},\mathrm{B};\boldsymbol{\gamma}\}\frac{\partial}{\partial \boldsymbol{\beta}} log\{f(\mathrm{Y}|\mathbf{X};\boldsymbol{\beta}\}]\big] \\
\mathbf{L} &= \mathrm{E_{XB}}\big[\mathrm{u}_\gamma(\mathbf{X},\mathrm{B})\mathrm{u}_\gamma^{\mathrm{T}}(\mathbf{X},\mathrm{B})\big]
\end{aligned}$$

with $\mathrm{u}_\gamma(\mathbf{X},\mathrm{B}) = \mathrm{E_{Y|XB}}\big[\frac{\partial}{\partial \boldsymbol{\beta}} log\{f(\mathrm{Y}|\mathbf{X};\boldsymbol{\beta})\}\big]$.

Intuitively, the asymptotic variance of this estimator is the inverse of information matrix $\mathbf{I}$ of $f(\mathrm{Y}|\mathbf{X},\mathrm{B};\boldsymbol{\gamma})$ plus the additional information due to knowing $\boldsymbol{\beta}$ from the external study $\mathbf{C}\mathbf{L}^{-1}\mathbf{C}^{\mathrm{T}}$.

2.5.2 Description of Two Special Cases

In the following two special cases, the goal is to derive the asymptotic efficiency of $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_{\mathrm{X}}, \hat{\gamma}_{\mathrm{B}})^{\mathrm{T}}$ through a closed-form expression for $\mathrm{Var}(\hat{\boldsymbol{\gamma}})$, and then compare the efficiency gain among all three approaches through the Asymptotic Relative Efficiency (ARE) of Var($\hat{\boldsymbol{\gamma}}$)'s, compared to Var($\hat{\boldsymbol{\gamma}}$) from the standard MLE. This will show how much efficiency we can gain by incorporating the external information from the external $\mathrm{Y}|\mathbf{X}$ model and what determines that gain. We describe the settings in this section and summarize the results in the subsequent section. For special case 1, the detailed algebraic derivation for each of the three approaches can be found in Appendix A.1.2, A.1.3 and A.1.4, respectively. Similarly for special case 2, the detailed algebraic derivation can be found in Appendix A.2.1, A.2.2 and A.2.3.

2.5.2.1 Special Case 1: Y and B are Gaussian Distributed

In this section, we assume that Y and B are continuous and have a Gaussian distribution, and assume the identity link for g_1 and g_2 in models 2.3 and 2.4.

Without loss of generality, we consider a simplified situation where there is only one X, i.e. p = 1. We also assume zero marginal means for Y, X and B, thus we will have no-intercept models as shown below. Let σ^2_{X} denote the variance of X. Then

$$\mathrm{Y}|\mathrm{X} \sim \mathrm{N}(\beta\mathrm{X}, \sigma^2_\beta) \tag{2.5}$$

$$\mathrm{Y}|\mathrm{X}, \mathrm{B} \sim \mathrm{N}(\gamma_{\mathrm{X}}\mathrm{X} + \gamma_{\mathrm{B}}\mathrm{B}, \sigma^2_\gamma) \tag{2.6}$$

$$\mathrm{B}|\mathrm{X} \sim \mathrm{N}(\theta\mathrm{X}, \sigma^2_\theta) \tag{2.7}$$

Depending on the information available from the external model $\mathrm{Y}|\mathrm{X}$, we consider two possible situations which correspond to two different constraints. The first situation is when the estimated coefficient $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ is known from model 2.5. This gives us the constraint $\theta = \theta^* = \frac{\beta^* - \gamma_{\mathrm{X}}}{\gamma_{\mathrm{B}}}$. The second situation is when both of the estimated coefficient $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ and the variance $\sigma^2_\beta = \sigma^{*2}_\beta = \gamma^2_{\mathrm{B}}\sigma^{*2}_\theta + \sigma^2_\gamma$ are known.

For the standard MLE of the complete dataset of size n, it is easy to show that the asymptotic variance of $\hat{\gamma}_{\mathrm{X}}$ and $\hat{\gamma}_{\mathrm{B}}$ are equal to $\frac{\sigma^2_\gamma}{\mathrm{n}\sigma^2_\theta}(\theta^2 + \frac{\sigma^2_\theta}{\sigma^2_{\mathrm{X}}})$ and $\frac{\sigma^2_\gamma}{\mathrm{n}\sigma^2_\theta}$, respectively.

2.5.2.2 Special Case 2: $\mathrm{Y}, \mathrm{X}, \mathrm{B}$ are All Binary

Assume we are interested in a saturated model:

$$\mathit{logit}[\Pr(\mathrm{Y} = 1|\mathrm{X}, \mathrm{B})] = \gamma_0 + \gamma_{\mathrm{X}}\mathrm{X} + \gamma_{\mathrm{B}}\mathrm{B} + \gamma_{\mathrm{XB}}\mathrm{XB} \tag{2.8}$$

describing the joint effect of X, B on Y, when $\mathrm{Y}, \mathrm{X}, \mathrm{B}$ are all binary variables. The external information from model 2.3 can be expressed as:

$$\mathit{logit}[\Pr(\mathrm{Y} = 1|\mathrm{X})] = \beta_0 + \beta_1\mathrm{X} \tag{2.9}$$

The association between B and X is defined through the model:

$$\mathit{logit}[\Pr(\mathrm{B} = 1|\mathrm{X})] = \theta_0 + \theta_1\mathrm{X}$$

We denote $\mathrm{P}(\mathrm{X} = \mathrm{a}, \mathrm{Y} = \mathrm{b})$ as the probability of $(\mathrm{X} = \mathrm{a}, \mathrm{Y} = \mathrm{b})$ combination and $\mathrm{P}(\mathrm{B} = 0, \mathrm{X} = \mathrm{a}, \mathrm{Y} = \mathrm{b})$ as the probability of $(\mathrm{B} = 0, \mathrm{X} = \mathrm{a}, \mathrm{Y} = \mathrm{b})$ combination, where $\mathrm{a}, \mathrm{b} \in \{0, 1\}$.

2.5.3 Summary

Based on the derivation listed in Appendix A.1 and A.2, we summarize the methods and the assumed forms of the summary-level external information for each approach in Tables 2.6 and 2.7 for special case 1 and 2, respectively.

Table 2.6: Summary of the special case 1 (Y and B are Gaussian)

Approach	Method for including external information	Available form of the external information	ARE$(\hat{\gamma})^*$ $\hat{\gamma}_{\mathrm{X}}$	$\hat{\gamma}_{\mathrm{B}}$
Standard MLE (ref)	None	NA	1	1
1: Synthetic data method	m additional synthetic data observations	Ability to draw Y values from Y|X distribution, regardless of the form	$1 - \mathrm{A}^{\dagger} - \frac{\sigma_{\mathrm{X}}^2\theta^{*2}}{\sigma_{\mathrm{X}}^2\theta^{*2}+\sigma_{\theta}^{*2}}\,\mathrm{D}^{\ddagger}$	$1 - \mathrm{D}$
2: Constrained MLE Cheng et al. [2018]	Constraint	2.1: The estimated coefficient β is known.	$1 - \mathrm{A}$	1
		2.2: Both of the estimated coefficient β and the standard deviation σ_β are known.	$1 - \mathrm{A} - \frac{\sigma_{\mathrm{X}}^2\theta^{*2}}{\sigma_{\mathrm{X}}^2\theta^{*2}+\sigma_{\theta}^{*2}}\mathrm{D}$	$1 - \mathrm{D}$
3: CSPML Chatterjee et al. [2016]	Constraint	Known expectation of Y|X	$1 - \mathrm{A}$	1

* ARE$(\hat{\gamma})$ = $\mathrm{Var_M}(\hat{\gamma})/\mathrm{Var_{MLE}}(\hat{\gamma})$, $\mathrm{M} \in \{\text{Synthetic Data}, \mathrm{CML}, \mathrm{CSPML}\}$

† A = $\frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_{\mathrm{X}}^2\theta^{*2}}\frac{\sigma_\gamma^2}{\sigma_\gamma^2+\gamma_{\mathrm{B}}^2\sigma_\theta^2}$, where $\theta^* = \frac{\beta^*-\gamma_{\mathrm{X}}}{\gamma_{\mathrm{B}}}$.

‡ D = $\frac{2\sigma_\gamma^2\gamma_{\mathrm{B}}^2\sigma_\theta^{*2}}{\sigma_\beta^{*4}}$, where $\sigma_\theta^{*2} = \frac{\sigma_\beta^{*2}-\sigma_\gamma^2}{\gamma_{\mathrm{B}}^2}$.

Table 2.7: Summary of the special case 2 (Y, X and B are binary)

Approach	Method for including external information	Available form of the external information	ARE$(\hat{\gamma})^*$ $\hat{\gamma}_0$	$\hat{\gamma}_{\mathrm{X}}$	$\hat{\gamma}_{\mathrm{B}}$ ($\hat{\gamma}_{\mathrm{XB}}$)
Standard MLE (ref)	None	NA	1	1	1
1: Synthetic data method	m additional synthetic data observations	Ability to draw Y values from Y|X distribution, regardless of the form	$1 - \mathrm{F}^{\dagger}$	$1 - \mathrm{G}^{\ddagger}$	1
2: Constrained MLE Cheng et al. [2018]	Constraint	Known estimated coefficient β	$1 - \mathrm{F}$	$1 - \mathrm{G}$	1
3: CSPML Chatterjee et al. [2016]	Constraint	Known expectation of Y|X	$1 - \mathrm{F}$	$1 - \mathrm{G}$	1

* ARE = $\mathrm{Var_M}(\hat{\gamma})/\mathrm{Var_{MLE}}(\hat{\gamma})$, $\mathrm{M} \in \{\text{Synthetic Data}, \mathrm{CML}, \mathrm{CSPML}\}$

† F = $\frac{\sum_{(\mathrm{a,b})\in\{(0,1)(0,0)\}} 1/\mathrm{P(X=a,Y=b)}}{\sum_{(\mathrm{a,b})\in\{(0,1)(0,0)\}} 1/\mathrm{P(B=0,X=a,Y=b)}}$. ‡ G = $\frac{\sum_{\mathrm{a,b}\in\{0,1\}} 1/\mathrm{P(X=a,Y=b)}}{\sum_{\mathrm{a,b}\in\{0,1\}} 1/\mathrm{P(B=0,X=a,Y=b)}}$.

Let $\mathrm{ARE_M}(\hat{\gamma}) = \frac{\mathrm{Var_M}(\hat{\gamma})}{\mathrm{Var_{MLE}}(\hat{\gamma})}$ denotes the asymptotic relative efficiency under approach M relative to the standard MLE (without external information), where $\mathrm{M} \in \{\text{Synthetic Data}, \mathrm{CML(2.1)}, \mathrm{CML(2.2)}, \mathrm{CSPML}\}$ for Gaussian Y, and $\mathrm{M} \in \{\text{Synthetic Data}, \mathrm{CML}, \mathrm{CSPML}\}$ for binary Y.

Result 1. Special case 1: Y and B are continuous and normally distributed (Table 2.6)

- $\mathrm{ARE_{Synthetic\ Data}}(\hat{\gamma}_\mathrm{X}) = \mathrm{ARE_{CML(2.2)}}(\hat{\gamma}_\mathrm{X}) = 1 - \mathrm{A} - \frac{\sigma_\mathrm{X}^2\theta^{*2}}{\sigma_\mathrm{X}^2\theta^{*2}+\sigma_\theta^{*2}}\mathrm{D}$
  $\mathrm{ARE_{CSPML}}(\hat{\gamma}_\mathrm{X}) = \mathrm{ARE_{CML(2.1)}}(\hat{\gamma}_\mathrm{X}) = 1 - \mathrm{A}$
- $\mathrm{ARE_{Synthetic\ Data}}(\hat{\gamma}_\mathrm{B}) = \mathrm{ARE_{CML(2.2)}}(\hat{\gamma}_\mathrm{B}) = 1 - \mathrm{D}$
  $\mathrm{ARE_{CSPML}}(\hat{\gamma}_\mathrm{B}) = \mathrm{ARE_{CML(2.1)}}(\hat{\gamma}_\mathrm{B}) = 1$

where $\mathrm{A} = \frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_\mathrm{X}^2\theta^{*2}}\frac{\sigma_\gamma^2}{\sigma_\gamma^2+\gamma_\mathrm{B}^2\sigma_\theta^2}$, $\mathrm{D} = \frac{2\sigma_\gamma^2\gamma_\mathrm{B}^2\sigma_\theta^{*2}}{\sigma_\beta^{*4}}$, $\theta^* = \frac{\beta^*-\gamma_\mathrm{X}}{\gamma_\mathrm{B}}$ and $\sigma_\theta^{*2} = \frac{\sigma_\beta^{*2}-\sigma_\gamma^2}{\gamma_\mathrm{B}^2}$. In summary, the synthetic data method has the same asymptotic variance as the CML (approach 2.2), and both are more efficient than the CSPML and the CML (approach 2.1). For $\hat{\gamma}_\mathrm{X}$ the CSPML and the CML (approach 2.1) are more efficient than the standard MLE. For $\hat{\gamma}_\mathrm{B}$ the CSPML and the CML (approach 2.1) have the same efficiency as the standard MLE.

Result 2. Special case 2: Y, X, and B are all binary (Table 2.7)

- $\mathrm{ARE_{Synthetic\ Data}}(\hat{\gamma_0}) = \mathrm{ARE_{CML}}(\hat{\gamma}_0) = \mathrm{ARE_{CSPML}}(\hat{\gamma}_0) = 1 - \mathrm{F}$
- $\mathrm{ARE_{Synthetic\ Data}}(\hat{\gamma}_\mathrm{X}) = \mathrm{ARE_{CML}}(\hat{\gamma}_\mathrm{X}) = \mathrm{ARE_{CSPML}}(\hat{\gamma}_\mathrm{X}) = 1 - \mathrm{G}$
- $\mathrm{ARE_{Synthetic\ Data}}(\hat{\gamma}_\mathrm{B}) = \mathrm{ARE_{CML}}(\hat{\gamma}_\mathrm{B}) = \mathrm{ARE_{CSPML}}(\hat{\gamma}_\mathrm{B}) = 1$
- $\mathrm{ARE_{Synthetic\ Data}}(\hat{\gamma}_\mathrm{XB}) = \mathrm{ARE_{CML}}(\hat{\gamma}_\mathrm{XB}) = \mathrm{ARE_{CSPML}}(\hat{\gamma}_\mathrm{XB}) = 1$

where $\mathrm{F} = \frac{\sum_{(\mathrm{a,b})\in\{(0,1)(0,0)\}} 1/\mathrm{P(X=a,Y=b)}}{\sum_{(\mathrm{a,b})\in\{(0,1)(0,0)\}} 1/\mathrm{P(B=0,X=a,Y=b)}}$ and $\mathrm{G} = \frac{\sum_{\mathrm{a,b}\in\{0,1\}} 1/\mathrm{P(X=a,Y=b)}}{\sum_{\mathrm{a,b}\in\{0,1\}} 1/\mathrm{P(B=0,X=a,Y=b)}}$. In conclusion, the synthetic data method, CML and CSPML all converge to the same asymptotic variance. We notice that $\hat{\gamma}_0$ and $\hat{\gamma}_\mathrm{X}$ are more efficient than the standard MLE while $\hat{\gamma}_\mathrm{B}$ and $\hat{\gamma}_\mathrm{XB}$ have the same efficiency as the standard MLE.

2.5.4 Justification from Another Perspective

In the two special cases, we show that using the synthetic data approach with very large m (i.e. total size of synthetic data) gives identical asymptotic variance for the parameters of model 2.1 as the constrained ML approach. Below we provide a different intuitive justification for the synthetic data approach, for a more general situation, if certain conditions apply. Assume that Y

and B are scalar random variables and that $\mathbf{X}$ is a vector of covariates. We will assume parametric models for all the conditional distributions, and these can be written as $f(\mathrm{Y},\mathrm{B}|\mathbf{X},\boldsymbol{\phi})$, $f(\mathrm{Y}|\mathbf{X},\mathrm{B};\boldsymbol{\gamma})$, $f(\mathrm{Y}|\mathbf{X};\boldsymbol{\beta})$, $f(\mathrm{B}|\mathbf{X};\theta)$ and $f(\mathrm{B}|\mathbf{X},\mathrm{Y};\boldsymbol{\kappa})$. Assume that $f(\mathrm{Y}|\mathbf{X},\mathrm{B};\boldsymbol{\gamma})$ is the model of interest, and that $f(\mathrm{Y}|\mathbf{X};\boldsymbol{\beta})$ is the form of the model that was fit to the external data, and that the estimate of $\boldsymbol{\beta}$ from the external data approximates the true value of $\boldsymbol{\beta}$. We assume that all these models represent the true distributions and are compatible with each other in the sense that $f(\mathrm{Y},\mathrm{B}|\mathbf{X},\boldsymbol{\phi}) = f(\mathrm{Y}|\mathbf{X},\mathrm{B};\boldsymbol{\gamma}) \times f(\mathrm{B}|\mathbf{X};\boldsymbol{\theta}) = f(\mathrm{B}|\mathbf{X},\mathrm{Y};\boldsymbol{\kappa}) \times f(\mathrm{Y}|\mathbf{X};\boldsymbol{\beta})$. We assume there is a 1-to-1 mapping between $\boldsymbol{\phi}$ and $(\boldsymbol{\gamma},\boldsymbol{\theta})$ and between $\boldsymbol{\phi}$ and $(\boldsymbol{\kappa},\boldsymbol{\beta})$, and that $\boldsymbol{\kappa}$ and $\boldsymbol{\beta}$ are distinct and that $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ are distinct. With these conditions, we can write $f(\mathrm{Y},\mathrm{B}|\mathbf{X};\boldsymbol{\phi})$ as $f(\mathrm{Y},\mathrm{B}|\mathbf{X};\boldsymbol{\kappa},\boldsymbol{\beta})$.

With this set-up, the CML estimate is obtained by maximizing the likelihood $\prod_{\mathrm{i}=1}^{\mathrm{n}} f(\mathrm{Y_i},\mathrm{B_i}|\mathbf{X_i};\boldsymbol{\phi})$ over $\boldsymbol{\phi}$, subject to the known $\boldsymbol{\beta}$. This can be rewritten as maximizing the likelihood $\prod_{\mathrm{i}=1}^{\mathrm{n}} f(\mathrm{Y_i},\mathrm{B_i}|\mathbf{X_i};\boldsymbol{\kappa},\boldsymbol{\beta})$ over $\boldsymbol{\kappa}$, subject to the known $\boldsymbol{\beta}$. Then from the combination of the estimate of $\boldsymbol{\kappa}$ and the known $\boldsymbol{\beta}$ we can obtain the estimate of $\boldsymbol{\gamma}$.

The synthetic data method consists of maximizing the likelihood

$$\prod_{\mathrm{i}=1}^{\mathrm{n}} f(\mathrm{Y_i},\mathrm{B_i}|\mathbf{X_i};\boldsymbol{\phi}) \prod_{\mathrm{i}=\mathrm{n}+1}^{\mathrm{n}+\mathrm{m}} f(\mathrm{Y_i}|\mathbf{X_i};\boldsymbol{\beta})$$

which is equivalent to maximizing

$$\prod_{\mathrm{i}=1}^{\mathrm{n}} f(\mathrm{Y_i},\mathrm{B_i}|\mathbf{X_i};\boldsymbol{\kappa},\boldsymbol{\beta}) \prod_{\mathrm{i}=\mathrm{n}+1}^{\mathrm{n}+\mathrm{m}} f(\mathrm{Y_i}|\mathbf{X_i};\boldsymbol{\beta})$$

over $\boldsymbol{\kappa}$ and $\boldsymbol{\beta}$. When optimizing over $\boldsymbol{\beta}$ for fixed $\boldsymbol{\kappa}$, the second term $\prod_{\mathrm{i}=\mathrm{n}+1}^{\mathrm{n}+\mathrm{m}} f(\mathrm{Y_i}|\mathbf{X_i};\boldsymbol{\beta})$ will dominate the optimization procedure when m is very large. Thus the estimate of $\boldsymbol{\beta}$ will essentially reproduce the known value from the external data (since this was the value used to generate the synthetic data). Thus the synthetic data method will reduce to the maximization of the remaining part of the likelihood $\prod_{\mathrm{i}=1}^{\mathrm{n}} f(\mathrm{Y_i},\mathrm{B_i}|\mathbf{X_i};\boldsymbol{\kappa},\boldsymbol{\beta})$ with $\boldsymbol{\beta}$ fixed, which is exactly the CML method.

The requirement that all the conditional distributions are compatible with each other will not usually be true, but maybe a reasonable approximation if flexible enough models are being used. The conditions do hold for the normal and the tri-binary examples in Section 2.5.2.1 and 2.5.2.2. Another case where they hold is when Y and B follow a bivariate normal distribution given $\mathbf{X}$, i.e. $\mathrm{Y},\mathrm{B}|\mathbf{X} \sim \mathrm{N}\left(\begin{pmatrix}\boldsymbol{\beta}\mathbf{X}\\ \boldsymbol{\theta}\mathbf{X}\end{pmatrix}, \begin{pmatrix}\sigma_\beta^2 & \rho\sigma_\beta\sigma_\theta\\ \rho\sigma_\beta\sigma_\theta & \sigma_\theta^2\end{pmatrix}\right)$. Then the CML is to maximize the likelihood $\prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{f}(\mathrm{Y_i},\mathrm{B_i}|\mathbf{X_i};\boldsymbol{\beta},\boldsymbol{\theta},\sigma_\beta,\sigma_\theta,\rho)$ over $\boldsymbol{\theta}$, σ_θ and ρ subject to the known $\boldsymbol{\beta}$ and σ_β.

2.6 Discussion

In this chapter, we introduced the synthetic data method for incorporating summary-level information from well-established external models into the regression model estimation based on internal data. We demonstrate in some special cases that with a large number of synthetic data observations, the synthetic data approach is asymptotically as efficient as the CML approach. This provides some justification for what at first sight might seem to be an ad-hoc approach. In a simulation study, we demonstrate the ability of the method to improve the predictive ability of the model.

A key advantage of the synthetic data method is that it naturally incorporates the prior knowledge into the internal data by creating large "fake" data that is compatible with the $\mathrm{Y}|\mathbf{X}$ established model. By creating pseudo-data from $\mathrm{Y}|\mathbf{X}$ instead of constrained optimization, the synthetic data method not only simplified the task from solving complex constrained optimization, but also provides a potentially more flexible and general framework to handle this problem. The only requirement for the synthetic data approach is the ability to generate Y values given $\mathbf{X}$ from the information of the external models, without the need to know the exact form of the model. It is broadly applicable for general data types for $\mathrm{Y}, \mathbf{X}$ and B, and when B is more than one new biomarker. It can be extended to the situation where more than one external model is available, i.e. $\mathrm{Y}|\mathbf{X_1}$, $\mathrm{Y}|\mathbf{X_2}$,...,$\mathrm{Y}|\mathbf{X_k}$. In this setting, a combination of external studies that measured overlapping but necessarily identical covariates can provide joint information to develop a model for $\mathrm{Y}|\mathbf{X}$ model, where $\mathbf{X}$ is the union of $\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_k}$. We will elaborate this in Chapter 4.

The CSPML approach is also broadly applicable, and can handle multiple B's, and it has some optimality properties. But it does require knowledge of the form of the $\mathrm{Y}|\mathbf{X}$ model and requires that the distribution of the X's are identical in the external and the internal populations, which seems unlikely to be satisfied in practice. When analyzing the synthetic dataset, the value of B can be considered as missing, which converts the problem of incorporating external information into a problem of analyzing data with missing values. If multiple imputation procedures are to be used to impute the value of B, then further research would be needed to suggest efficient and robust ways in which this should be implemented. There is the potential to improve even further on the method by using different ways of imputing B, beyond the approach we illustrated in the simulation study.

Another interesting issue that will need to be investigated is the size of m. The theoretical result in this chapter suggests that m should be very large, but this is under the assumption that the $\mathrm{Y}|\mathbf{X}$ and $\mathrm{Y}|\mathbf{X}, \mathrm{B}$ models are compatible with each other. In practice, they are unlikely to be exactly compatible, which would suggest limiting the size of m. A pragmatic suggestion is to make m equal to the size of the external data, if that is known. By doing this the amount of information in the synthetic data about the relationship between Y and $\mathbf{X}$ is similar to the amount of information in the external data about the relationship between Y and $\mathbf{X}$.

2.7 Publication

The content of this chapter has been published in *Canadian Journal of Statistics* at DOI: 10.1002/cjs.11513.

CHAPTER 3

A Meta-Inference Framework to Integrate Multiple External Models into a Current Study

3.1 Introduction

In Chapter 2, we introduced a synthetic data method as a flexible alternative to the constrained maximum likelihood (CML) approach. In two special cases and extensive simulation studies, we showed that the synthetic data method had the same asymptotic properties as a constrained semi-parametric maximum likelihood (CSPML) approach [Chatterjee et al., 2016], who converted the external summary-level information into a constraint and then maximized the internal data likelihood subject to this constraint.

Several methods were built upon the work of Chatterjee et al. [2016], e.g. Zhang et al. [2020] extended CSPML to account for parameter uncertainty of the external model while Kundu, Tang, and Chatterjee [2019] extended CSPML to a generalized meta-analysis approach. However, the CSPML method requires the joint distribution of (Y, X, B) to be the same in the internal and the external population, a strong assumption, which although unverifiable, we expect would be frequently violated, and can cause bias when violated. Estes, Mukherjee, and Taylor [2017] later proposed a matrix-weighted average remedy by constructing an empirical Bayes (EB) estimator that can reduce the potential bias. As an extension and adaption of Estes et al. [2017], in this chapter, we propose a meta-inference framework using a composite of EB estimators to accommodate the situation where multiple external prediction models are available to help improve the inference of the current study.

We consider the situation in which there are K external studies ($K \geq 2$), each of which developed a prediction model for the same outcome. The parameter estimates of the external models are known, but the individual-level data are not available. The goal is to develop a prediction model that uses all the possible covariates, using data from an internal study and the parameter estimates from the external models. The parameters of this model are the quantities of interest. Each of the external studies may use a slightly different set of covariates but the internal data are assumed

to contain all available covariates, as well as the new biomarkers that are not included in any of the external models. We propose a meta-inference framework using an empirical Bayes estimation approach, which first separately incorporates the different summary information from each external study into the internal study, and then takes a weighted average of the resulting estimators to give a final overall estimate of the parameters of interest. We show that the proposed final estimators are more efficient than the simple analysis of the internal data, as well as outperform the estimators that integrate the information from a single external model.

The rest of this chapter is organized as follows: In Section 3.2, we introduce two key existing estimators before introducing the proposed methodology along with two weighted estimators, followed by corresponding large sample results. In Section 3.3, we demonstrate the potential of our proposed method through a simulation study. In Section 3.4, we apply the proposed method to predict the risk of high-grade prostate cancer incorporating information from two existing risk prediction models. We present a discussion in Section 3.5.

3.2 Models and Methods

3.2.1 General Description of the Problem

Let Y denote the outcome of interest, which can be either continuous or binary. $\mathbf{X}$ is a set of p standard variables and let B denote a new biomarker. Our target of interest is the mean structure of $\mathrm{Y}|\mathbf{X}, \mathrm{B}$:

$$g(\mathrm{E}(\mathrm{Y}|\mathbf{X}, \mathrm{B})) = \mathbf{X}\boldsymbol{\gamma}_{\mathrm{X}} + \mathrm{B}\gamma_{\mathrm{B}} = \gamma_{\mathrm{X}_0} + \gamma_{\mathrm{X}_1}\mathrm{X}_1 + ... + \gamma_{\mathrm{X}_\mathrm{p}}\mathrm{X}_\mathrm{p} + \gamma_{\mathrm{B}}\mathrm{B}, \tag{3.1}$$

where g is the known link function. We assume that a small dataset of size n with variables $\mathrm{Y}, \mathbf{X}$ and a new covariate B is available to us for building model 3.2.1. For each external study k, k $\in \{2, ..., \mathrm{K}\}$, a prediction model for the same outcome Y has been built using predictors $\mathbf{X}_\mathrm{k}$, a subset of the internal $\mathbf{X}$. Each external model may use slightly different predictors to predict Y:

$$g(\mathrm{E}(\mathrm{Y}|\mathbf{X}_\mathrm{k})) = \mathbf{X}_\mathrm{k}\boldsymbol{\beta}_\mathrm{k} = \beta_0 + \beta_1\mathrm{X}_1 + \cdots + \beta_{\mathrm{p}_\mathrm{k}}\mathrm{X}_{\mathrm{p}_\mathrm{k}},$$

where $\mathrm{p}_\mathrm{k} \leq \mathrm{p}$ is the dimension of $\mathbf{X}_\mathrm{k}$. We assume that the distribution of $\mathrm{Y}|\mathbf{X}, \mathrm{B}$ is correctly specified, but the external $\mathrm{Y}|\mathbf{X}_\mathrm{k}$ distributions need not be.

We assume K large, well-characterized previous studies from the external populations describe the provided information on the calculated distribution of $\mathrm{Y}|\mathbf{X}_\mathrm{k}$. These information will come in the forms of estimated model parameters $\hat{\boldsymbol{\beta}}_\mathrm{k}$. The goal is to develop a framework, in which we can utilize all K external $\hat{\boldsymbol{\beta}}_\mathrm{k}$'s to improve the estimation efficiency of the internal study.

We introduce some of the important notation that will frequently appear in later sections:

- $\mathrm{f}_\beta(\mathrm{Y}|\mathrm{X}_\mathrm{k})$: the study-specific distribution of the k^{th} external model $\mathrm{Y}|\mathbf{X}_\mathrm{k}$;
- $\mathrm{f}_\gamma(\mathrm{Y}|\mathrm{X},\mathrm{B})$: distribution of the target model $\mathrm{Y}|\mathbf{X},\mathrm{B}$;
- $\hat{\boldsymbol{\gamma}}_\mathrm{I}$: the unconstrained estimator by using the internal data only;
- $\hat{\boldsymbol{\gamma}}_{\mathrm{CSPML}}$: the constrained semi-parametric maximum likelihood estimator (CSPML) proposed by Chatterjee et al. [2016];
- $\hat{\boldsymbol{\gamma}}_{\mathrm{EB}}$: the empirical Bayes (EB) estimator proposed by Estes et al. [2017].

3.2.2 Two Existing Estimators

The proposed method was developed on the foundation of two existing methods, the CSPML approach [Chatterjee et al., 2016] and the EB approach [Estes et al., 2017]. The CSPML estimator considered the same problem described here in a special case where K=1; and EB estimator applied the empirical Bayes method to CSPML, calibrating the potential bias due to non-transportability. Therefore, it is necessary to introduce these two core methods first before considering the $\mathrm{K} > 1$ situation.

In the CSPML, the proposed estimator $\hat{\boldsymbol{\gamma}}_{\mathrm{CSPML}}$ incorporates the external regression coefficients to calibrate the current regression model. Denote $\mathrm{U}_\beta(\mathrm{Y}|\mathrm{X})$ as the score function of the external $\mathrm{Y}|\mathrm{X};\beta$ model. It converts the external model parameter $\hat{\boldsymbol{\beta}}$ to a constraint by connecting the external score function with the target distribution $\mathrm{f}_\gamma(\mathrm{Y}|\mathrm{X},\mathrm{B})$:

$$\begin{aligned}
0 = \mathrm{E}_{\mathrm{Y,X,B}}[\mathrm{U}_\beta(\mathrm{Y}|\mathrm{X})] &= \mathrm{E}_{\mathrm{X,B}}\{\mathrm{E}_{\mathrm{Y|X,B}}[\mathrm{U}_\beta(\mathrm{Y}|\mathrm{X})|\mathrm{X},\mathrm{B}]\} \\
&= \int_{\mathrm{X,B}} \int_{\mathrm{Y|X,B}} \mathrm{U}_\beta(\mathrm{Y}|\mathrm{X})\mathrm{f}_\gamma(\mathrm{Y}|\mathrm{X},\mathrm{B})\mathrm{dY}\mathrm{dF}(\mathrm{X},\mathrm{B}) \\
&= \sum_{\mathrm{i}=1}^{\mathrm{n}} \int_{\mathrm{Y|X,B}} \mathrm{U}_\beta(\mathrm{Y}|\mathrm{X})\mathrm{f}_\gamma(\mathrm{Y}|\mathrm{X},\mathrm{B})\mathrm{dY}\mathrm{p}_\mathrm{i},
\end{aligned}$$

where $\mathrm{dF}(\mathrm{X},\mathrm{B})$ is the empirical probability distribution $\mathrm{p}_\mathrm{i} = \Pr(\mathrm{X}=\mathrm{X}_\mathrm{i}, \mathrm{B}=\mathrm{B}_\mathrm{i})$ for the internal observations and $\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{p}_\mathrm{i} = 1$. Then $\hat{\boldsymbol{\gamma}}_{\mathrm{CSPML}}$ was obtained using Lagrange multipliers by solving the following Lagrangian function:

$$\hat{\boldsymbol{\gamma}}_{\mathrm{CSPML}} = \arg\max_{\gamma,\mathrm{p}_\mathrm{i}} \{\prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{f}_\gamma(\mathrm{Y}_\mathrm{i}|\mathrm{X}_\mathrm{i},\mathrm{B}_\mathrm{i})\mathrm{p}_\mathrm{i} + \lambda_1(\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{p}_\mathrm{i} - 1) + \lambda_2 \sum_{\mathrm{i}=1}^{\mathrm{n}} \int \mathrm{U}_\beta(\mathrm{Y}|\mathrm{X})\mathrm{f}_\gamma(\mathrm{Y}|\mathrm{X},\mathrm{B})\mathrm{dY}\mathrm{p}_\mathrm{i}\}$$

Chatterjee et al. [2016] provided the asymptotic variance of $\hat{\gamma}_{\mathrm{CSPML}}$, showing the efficiency gain of $\hat{\gamma}_{\mathrm{CSPML}}$ compared to $\hat{\gamma}_{\mathrm{I}}$.

Estes et al. [2017] showed that in the CSPML approach, the strict assumption of the identical joint distribution of (Y,X,B) between the internal and the external population (also known as the full transportable assumption of the joint distribution of [Y,X,B]) is hard to satisfy in reality, and ignoring it can lead to substantial bias. Assuming the target conditional distribution $(\mathrm{Y}|\mathrm{X},\mathrm{B};\gamma)$ is correctly specified in the internal study and the underlying true parameter γ follows a stochastic framework, Estes et al. [2017] proposed an empirical Bayes (EB) estimator $\hat{\gamma}_{\mathrm{EB}}$, which can be viewed as a matrix-weighted average of the internal estimate $\hat{\gamma}_{\mathrm{I}}$ and the CSPML estimator $\hat{\gamma}_{\mathrm{CSPML}}$. The EB estimator uses the difference $\hat{\gamma}_{\mathrm{I}} - \hat{\gamma}_{\mathrm{CSPML}}$ to measure the distributional similarity of the joint distribution (Y, X, B) between the internal and the external population, and will down-weight $\hat{\gamma}_{\mathrm{CSPML}}$ if the lack of full transportability leads to a poor estimate. Therefore, the EB estimator is robust to departures from full transportability assumption in specific external populations. The features and assumptions for the CSPML, the EB and the proposed estimators are summarized in Appendix B.1.

Specifically, the EB approach first posits a stochastic framework connecting the internal estimator $\hat{\gamma}_{\mathrm{I}}$ and the underlying true parameter $\boldsymbol{\gamma} \sim \mathrm{N}(\boldsymbol{\gamma}_0, \mathbf{A})$ for some covariance matrix $\mathbf{A}$. Since $\begin{cases} \boldsymbol{\gamma} \sim \mathrm{N}(\boldsymbol{\gamma}_0, \mathbf{A}) \\ \hat{\gamma}_{\mathrm{I}} | \boldsymbol{\gamma}\ \mathrm{N}(\boldsymbol{\gamma}, \boldsymbol{\Sigma}) \end{cases}$, the posterior Bayes estimate of $\boldsymbol{\gamma}$ equals to $\mathbf{A}(\boldsymbol{\Sigma}+\mathbf{A})^{-1}\hat{\gamma}_{\mathrm{I}} + \boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\mathbf{A})^{-1}\boldsymbol{\gamma}_0$. Replacing $\boldsymbol{\gamma}_0$ with the CSPML estimator $\hat{\gamma}_{\mathrm{CSPML}}$ and empirically estimating $\mathbf{A}$ and $\boldsymbol{\Sigma}$, we obtain the EB estimator $\hat{\gamma}_{\mathrm{EB}} = \hat{\mathbf{A}}(\hat{\boldsymbol{\Sigma}} + \hat{\mathbf{A}})^{-1}\hat{\gamma}_{\mathrm{I}} + \hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Sigma}} + \hat{\mathbf{A}})^{-1}\hat{\gamma}_{\mathrm{CSPML}} \stackrel{\mathrm{def}}{=} \hat{\mathbf{W}}\hat{\gamma}_{\mathrm{I}} + (\mathrm{I} - \hat{\mathbf{W}})\hat{\gamma}_{\mathrm{CSPML}}$, where $\hat{\mathbf{A}} = (\hat{\gamma}_{\mathrm{I}} - \hat{\gamma}_{\mathrm{CSPML}})(\hat{\gamma}_{\mathrm{I}} - \hat{\gamma}_{\mathrm{CSPML}})^{\mathrm{T}}$ quantifies the difference between $\hat{\gamma}_{\mathrm{I}}$ and $\hat{\gamma}_{\mathrm{CSPML}}$, $\hat{\boldsymbol{\Sigma}}$ is the MLE of the variance of $\hat{\gamma}_{\mathrm{I}}$, and $\hat{\mathbf{W}} = \hat{\mathbf{A}}(\hat{\boldsymbol{\Sigma}} + \hat{\mathbf{A}})^{-1}$ is the empirical weights. Therefore, the EB estimator can be viewed as a matrix generalization of a weighted average of vectors $\hat{\gamma}_{\mathrm{I}}$ and $\hat{\gamma}_{\mathrm{CSPML}}$, of which the empirical weights will shrink the EB estimator towards $\hat{\gamma}_{\mathrm{I}}$ when $\hat{\gamma}_{\mathrm{CSPML}}$ is biased. The EB estimator's shrinkage effect limits the impact of external model information that is not compatible with the internal data and thus protects against the severe bias. Furthermore, when the joint distribution of (Y,X,B) is similar in the two populations, more precision will be gained by incorporating the external model.

3.2.3 Proposed Meta-Framework for Inference

We build upon the empirical Bayes work by Estes et al. [2017] and generalize it to accommodate the situation where we can combine multiple external model estimates into the internal study. Our proposed meta-inference framework allows each established model to have different dimensions. The framework consists of generating an EB estimator for each of the external $\hat{\boldsymbol{\beta}}_{\mathrm{k}}$ from the fitted

regression $\mathrm{Y}|\mathbf{X}_\mathrm{k}$, and then constructing the final estimates of the target through a weighted average, considering the correlation structure among all the EB estimators.

The proposed framework contains two steps:

- Step 1: For each of the K external model estimates $\hat{\boldsymbol{\beta}}_\mathrm{k}$, first apply the CSPML method [Chatterjee et al., 2016], and then apply the EB method [Estes et al., 2017]:

$$\text{Internal data + External } \hat{\boldsymbol{\beta}}_k \xrightarrow{\text{CSPML}} \hat{\boldsymbol{\gamma}}_{\mathrm{CSPML_k}} \xrightarrow{\text{EB}} \hat{\boldsymbol{\gamma}}_{\mathrm{EB_k}} = \hat{\mathrm{W}}_\mathrm{k}\hat{\boldsymbol{\gamma}}_\mathrm{I} + (\mathrm{I} - \hat{\mathrm{W}}_\mathrm{k})\hat{\boldsymbol{\gamma}}_{\mathrm{CSPML_k}},$$

  after which we obtain a total of K $\hat{\boldsymbol{\gamma}}_{\mathrm{EB}}$'s.

- Step 2: We propose two estimators of $\boldsymbol{\gamma}$ to be the weighted average of $\hat{\boldsymbol{\gamma}}_{\mathrm{EB}}$'s, of the form $\sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{w_k}\hat{\boldsymbol{\gamma}}_{\mathrm{EB_k}}$, in which each element of the final estimate is the weighted average of the K separate estimates for that element. One composite estimator is called the optimal covariance weighted estimator (OCWE) and another is called the selective coefficient learner (SC-Learner).

In step 1, we separately integrate each of the external $\hat{\boldsymbol{\beta}}$'s with the internal data, and the EB method accounts for the potential bias caused by the heterogeneity of the internal and that specific external population. This first step also unifies the disparate dimensions of the external models to be the same as the target model 3.1, and improves the efficiency of parameter estimates for those covariates that were used in the external models.

In step 2, the challenge is to combine K correlated vectors of EB estimators while maximizing the efficiency gain of the overall prediction. The simplest, yet not the most attractive, solution is to naively average $\hat{\boldsymbol{\gamma}}_{\mathrm{EB's}}$, i.e. $\frac{1}{\mathrm{K}}\sum_{\mathrm{k}=1}^{\mathrm{K}} \hat{\boldsymbol{\gamma}}_{\mathrm{EB_k}}$. Better weighting approaches take into account the variance and/or correlation among $\hat{\boldsymbol{\gamma}}_{\mathrm{EB's}}$. One option is to use the inverse of the prediction variances as weights, i.e. $\mathrm{w_k} = \frac{1/\sum_{\mathrm{i}=1}^{\mathrm{n}} \hat{\mathrm{V}}\mathrm{ar}[(\mathbf{X_i},\mathrm{B_i})\hat{\boldsymbol{\gamma}}_{\mathrm{EB_k}}]}{\sum_{\mathrm{k}=1}^{\mathrm{K}} 1/\sum_{\mathrm{i}=1}^{\mathrm{n}} \hat{\mathrm{V}}\mathrm{ar}[(\mathbf{X_i},\mathrm{B_i})\hat{\boldsymbol{\gamma}}_{\mathrm{EB_k}}]}$, with the same $\mathrm{w_k}$ used for all elements of $\boldsymbol{\gamma}$. This method incorporates the within-estimator variance while ignoring the between-estimator covariance (i.e. ignoring the fact that the $\hat{\boldsymbol{\gamma}}_{\mathrm{EB}}$'s are not independent). Other popular design criterions that seek the optimal $\hat{\boldsymbol{\gamma}}$ to minimize the variance-covariance matrix of $\boldsymbol{\gamma}$ include D-optimality that minimizes the determinant of the matrix, A-optimality that minimizes the trace of the matrix, I-optimality (also known as V- or IV- or Q-optimality) that minimize the average prediction variance and G-optimality that minimizes the maximum prediction variance [Goos and Jones, 2011]. Since all criterions had similar performance in this study, we consider an adaptive version of I-optimality which seeks $\hat{\boldsymbol{\gamma}}$ that minimizes the average variance of the predicted estimator $\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \hat{\mathrm{V}}\mathrm{ar}[(\mathbf{X_i}, \mathrm{B_i})\hat{\boldsymbol{\gamma}}]$. We propose two weighted estimators that accounts for both within and between variances among $\hat{\boldsymbol{\gamma}}_{\mathrm{EB}}$'s:

1. **The optimal covariance weighted estimator (OCWE)**: OCWE views each $\hat{\boldsymbol{\gamma}}_{\mathrm{EB}}$ as a whole and provides the same weight $\mathrm{w_k}$ to each covariate coefficient within $\hat{\boldsymbol{\gamma}}_{\mathrm{EB_k}}$ that minimizes the overall estimated prediction variance, i.e.

$$\hat{\boldsymbol{\gamma}}_{\mathrm{OCWE}} = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \hat{\mathrm{V}}\mathrm{ar}[(\mathbf{X_i}, \mathrm{B_i})\hat{\boldsymbol{\gamma}}(\mathbf{w})],$$

   where $\hat{\boldsymbol{\gamma}}(\mathbf{w}) = \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{w_k}\hat{\boldsymbol{\gamma}}_{\mathrm{EB_k}}$ and $\mathbf{w} = (\mathrm{w}_1, \ldots, \mathrm{w_K})^{\mathrm{T}}$ denotes the positive weights that add up to one.

2. **The selective coefficient learner (SC-Learner)**: Instead of seeking a fixed weight for each $\hat{\boldsymbol{\gamma}}_{\mathrm{EB}}$ as in OCWE, SC-Learner attempts to find a set of weights separately for each covariate coefficient (from intercept $\hat{\gamma}_0$ to slopes $\hat{\gamma}_{\mathrm{X}_1}, ...\hat{\gamma}_{\mathrm{X_p}}, \hat{\gamma}_{\mathrm{B}}$) that minimize the corresponding variance, and thus each coefficient in one $\hat{\boldsymbol{\gamma}}_{\mathrm{EB}}$ can have different weights. Let $\mathrm{E_j}$ denotes the index set of the external models that included the predictor $\mathrm{X_j}$. For each predictor $\mathrm{X_j}$ (j $\in$ 0,...,p), SC-Learner first selects $\hat{\gamma}_{\mathrm{X_j^k}}$ from $\hat{\boldsymbol{\gamma}}_{\mathrm{EB_k}}$ which used $\mathrm{X_j}$ as a predictor in the external model, and then uses the inverse variance as the weight $\mathrm{w_{kj}} = \frac{1/\hat{\mathrm{V}}\mathrm{ar}(\hat{\boldsymbol{\gamma}}_{\mathrm{X_j^k}})}{\sum_{\mathrm{k}\in\mathrm{E_j}} 1/\hat{\mathrm{V}}\mathrm{ar}(\hat{\boldsymbol{\gamma}}_{\mathrm{X_j^k}})}$. The final estimate of each $\gamma_{\mathrm{X_j}}$ is an inverse variance-weighted estimator using selective coefficients from $\hat{\boldsymbol{\gamma}}_{\mathrm{EB}}$'s: $\hat{\gamma}^*_{\mathrm{X_j}} = \sum_{\mathrm{k}\in\mathrm{E_j}} \mathrm{w_{kj}}\hat{\boldsymbol{\gamma}}_{\mathrm{X_j^k}}$. We will use $\hat{\gamma}_{\mathrm{B}}$ from the direct regression $\hat{\boldsymbol{\gamma}}_{\mathrm{I}}$ as the final estimate for the B variable, since B is only available from the internal data and no external models have used B as predictors. Thus,

$$\hat{\boldsymbol{\gamma}}_{\mathrm{SC-Learner}} = [\hat{\gamma}^*_{\mathrm{X}_0}, \hat{\gamma}^*_{\mathrm{X}_1}, ..., \hat{\gamma}^*_{\mathrm{X_p}}, \hat{\gamma}_{\mathrm{B}}]^{\mathrm{T}}$$

   To illustrate this method, consider a hypothetical example with three external models–Model 1 $\mathrm{Y}|\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3$, Model 2 $\mathrm{Y}|\mathrm{X}_1, \mathrm{X}_2$, and Model 3 $\mathrm{Y}|\mathrm{X}_1, \mathrm{X}_3$ available to build the target model $\mathrm{Y}|\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3, \mathrm{B}$ together with the internal data. When considering the final estimated coefficient of X_3, $\hat{\gamma}_{\mathrm{X}_3^1}$ and $\hat{\gamma}_{\mathrm{X}_3^3}$ from the external models 1 and 3 will be used, while $\hat{\gamma}_{\mathrm{X}_3^2}$ that did not add extra information to X_3 will be excluded.

In both proposed estimators, we use the asymptotic variance-covariance structure derived from the large sample theory to capture the correlation among $\hat{\boldsymbol{\gamma}}_{\mathrm{EB's}}$, which will be discussed in detail in Section 3.2.4.

3.2.4 Asymptotic Normality and Large Sample Results

The following proposition extends the asymptotic normality of the CSPML estimator [Chatterjee et al., 2016] to higher dimension, as well as showing the correlation structure between $\hat{\boldsymbol{\gamma}}_{\mathrm{CSPML}}$ and

$\hat{\gamma}_{\mathrm{I}}$.

Proposition 1. Let $\hat{\boldsymbol{\eta}} = (\hat{\gamma}^{\mathrm{T}}_{\mathrm{CSPML}_1}, \ldots, \hat{\gamma}^{\mathrm{T}}_{\mathrm{CSPML}_\mathrm{K}}, \hat{\gamma}^{\mathrm{T}}_{\mathrm{I}})^{\mathrm{T}}$, and $\boldsymbol{\eta}_0 = (\boldsymbol{\gamma}^{\mathrm{T}}_0, \ldots, \boldsymbol{\gamma}^{\mathrm{T}}_0, \boldsymbol{\gamma}^{\mathrm{T}}_0)^{\mathrm{T}}$ with $\boldsymbol{\gamma}_0$ the true value of $\boldsymbol{\gamma}_{\mathrm{CSPML}}$ and $\boldsymbol{\gamma}_{\mathrm{I}}$. Under regularity conditions described in Chatterjee et al. [2016], as the internal sample size $\mathrm{n} \to \infty$, $\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0)$ converges in distribution to a normal distribution with zero-mean and covariance matrix given by

$$\begin{pmatrix} (\mathrm{B}+\mathrm{C}_1^{\mathrm{T}}\mathrm{L}_{11}^{-1}\mathrm{C}_1)^{-1} & \cdots & \cdots & \cdots & \cdots & (\mathrm{B}+\mathrm{C}_1^{\mathrm{T}}\mathrm{L}_{11}^{-1}\mathrm{C}_1)^{-1} \\ & \ddots & & \vdots & & \vdots \\ & (\mathrm{B}+\mathrm{C}_\mathrm{j}^{\mathrm{T}}\mathrm{L}_{\mathrm{jj}}^{-1}\mathrm{C}_\mathrm{j})^{-1} & \cdots & \mathrm{Cov}(\hat{\gamma}_{\mathrm{CSPML}_\mathrm{j}}, \hat{\gamma}_{\mathrm{CSPML}_\mathrm{k}}) & \cdots & (\mathrm{B}+\mathrm{C}_\mathrm{j}^{\mathrm{T}}\mathrm{L}_{\mathrm{jj}}^{-1}\mathrm{C}_\mathrm{j})^{-1} \\ & & \ddots & \vdots & & \vdots \\ & & & (\mathrm{B}+\mathrm{C}_\mathrm{K}^{\mathrm{T}}\mathrm{L}_{\mathrm{KK}}^{-1}\mathrm{C}_\mathrm{K})^{-1} & \cdots & (\mathrm{B}+\mathrm{C}_\mathrm{K}^{\mathrm{T}}\mathrm{L}_{\mathrm{KK}}^{-1}\mathrm{C}_\mathrm{K})^{-1} \\ & & & & & \mathrm{B}^{-1} \end{pmatrix}, \tag{3.2}$$

where $\mathrm{B} = -\mathrm{E}\{\frac{\partial^2}{\partial\gamma^{\mathrm{T}}\gamma}\mathrm{logf}_\gamma(\mathrm{Y}|\mathrm{X},\mathrm{B})\}$, $\mathrm{u}_\gamma(\beta_\mathrm{j}) = \mathrm{E}_{\mathrm{Y}|\mathrm{X},\mathrm{B}}\{\frac{\partial}{\partial\beta_\mathrm{j}}\mathrm{logf}_{\beta_\mathrm{j}}(\mathrm{Y}|\mathrm{X})\}$, $\mathrm{C_j} = \mathrm{E}\{\frac{\partial}{\partial\gamma}\mathrm{u}_\gamma(\beta_\mathrm{j})\}$, $\mathrm{L_{jk}} = \mathrm{E}\{\mathrm{u}^{\mathrm{T}}_\gamma(\beta_\mathrm{j})\mathrm{u}_\gamma(\beta_\mathrm{k}))\}$, $\mathrm{Cov}(\hat{\gamma}_{\mathrm{CSPML}_\mathrm{j}}, \hat{\gamma}_{\mathrm{CSPML}_\mathrm{k}}) = (\mathrm{B}+\mathrm{C}_\mathrm{j}^{\mathrm{T}}\mathrm{L}_{\mathrm{jj}}^{-1}\mathrm{C}_\mathrm{j})^{-1}(\mathrm{B}+\mathrm{C_j}\mathrm{L}_{\mathrm{jj}}^{-1}\mathrm{L_{jk}}\mathrm{L}_{\mathrm{kk}}^{-1}\mathrm{C_k})(\mathrm{B}+\mathrm{C}_\mathrm{k}^{\mathrm{T}}\mathrm{L}_{\mathrm{kk}}^{-1}\mathrm{C}_\mathrm{k})^{-1}$, and $\mathrm{j},\mathrm{k} \in \{1,\ldots,\mathrm{K}\}$.

As shown in *Proposition 1,* as well as in Chatterjee et al. [2016], the asymptotic variance of $\hat{\gamma}_{\mathrm{CSPML}}$ will decrease from B^{-1} to $(\mathrm{B}+\mathrm{C}_\mathrm{k}^{\mathrm{T}}\mathrm{L}_{\mathrm{kk}}^{-1}\mathrm{C}_\mathrm{k})^{-1}$ after incorporating the k^{th} external model information. We further show additional two conclusions: (i) the covariance between $\hat{\gamma}_{\mathrm{CSPML}_\mathrm{k}}$ and $\hat{\gamma}_{\mathrm{I}}$ is equivalent to the variance of $\hat{\gamma}_{\mathrm{CSPML}_\mathrm{k}}$, i.e. $\mathrm{Cov}(\hat{\gamma}_{\mathrm{CSPML}_\mathrm{k}}, \hat{\gamma}_{\mathrm{I}}) = \mathrm{Var}(\hat{\gamma}_{\mathrm{CSPML}_\mathrm{k}})$; (ii) the covariance between any two $\hat{\gamma}_{\mathrm{CSPML}_\mathrm{j}}$ and $\hat{\gamma}_{\mathrm{CSPML}_\mathrm{k}}$ equals to $(\mathrm{B}+\mathrm{C_j}\mathrm{L}_{\mathrm{jj}}^{-1}\mathrm{L_{jk}}\mathrm{L}_{\mathrm{kk}}^{-1}\mathrm{C_k})$ multiplied by their variances, i.e. $\mathrm{Cov}(\hat{\gamma}_{\mathrm{CSPML}_\mathrm{j}}, \hat{\gamma}_{\mathrm{CSPML}_\mathrm{k}}) = (\mathrm{B}+\mathrm{C}_\mathrm{j}^{\mathrm{T}}\mathrm{L}_{\mathrm{jj}}^{-1}\mathrm{C}_\mathrm{j})^{-1}(\mathrm{B}+\mathrm{C_j}\mathrm{L}_{\mathrm{jj}}^{-1}\mathrm{L_{jk}}\mathrm{L}_{\mathrm{kk}}^{-1}\mathrm{C_k})(\mathrm{B}+\mathrm{C}_\mathrm{k}^{\mathrm{T}}\mathrm{L}_{\mathrm{kk}}^{-1}\mathrm{C}_\mathrm{k})^{-1}$. In Appendix B.2, we show the extension of *Proposition 1* when the uncertainty of the external $\hat{\boldsymbol{\beta}}$ is known. As expected, this modification makes a difference only when the uncertainty is large.

Proposition 2. Let $\mathrm{Z} \equiv \hat{\gamma}_{\mathrm{I}} - \hat{\gamma}_{\mathrm{CSPML}}$ and $\hat{\mathrm{V}}_{\mathrm{I}} \equiv \hat{\mathrm{Var}}(\hat{\gamma}_{\mathrm{I}})$. We can re-parameterize $\hat{\gamma}_{\mathrm{EB}}$ as a function of Z and $\hat{\gamma}_{\mathrm{CSPML}}$:

$$\hat{\gamma}_{\mathrm{EB}} = \hat{\gamma}_{\mathrm{CSPML}} + \mathrm{Z}(1 - \frac{1}{1+\mathrm{Z}^{\mathrm{T}}\hat{\mathrm{V}}_{\mathrm{I}}^{-1}\mathrm{Z}}),$$

where $\mathrm{Z}^{\mathrm{T}}\hat{\mathrm{V}}_{\mathrm{I}}^{-1}\mathrm{Z}$ is a scalar. Equivalently, $\hat{\gamma}_{\mathrm{EB}}$ can be written as a function of Z and $\hat{\gamma}_{\mathrm{I}}$, i.e. $\hat{\gamma}_{\mathrm{EB}} = \hat{\gamma}_{\mathrm{I}} - \mathrm{Z}\frac{1}{1+\mathrm{Z}^{\mathrm{T}}\hat{\mathrm{V}}_{\mathrm{I}}^{-1}\mathrm{Z}}$. The proof is listed in Appendix B.3.

In the proposed method, we use the asymptotic variance-covariance structure derived from *Proposition 1* and *2* to capture the correlation among $\hat{\gamma}_{\mathrm{EB}'\mathrm{s}}$. Under the assumption that the external population is representative of the target population of interest, the mean of Z converges to zero. In Appendix B.3, we show that Z and $\hat{\gamma}_{\mathrm{CSPML}}$ are independently normal-distributed with closed-form mean and variance, from which it is easy to simulate many values

of Z and $\hat{\gamma}_{\mathrm{CSPML}}$. Therefore, according to *Proposition 2* and denote f(Z) as $\mathrm{Z}(1-\frac{1}{1+\mathrm{Z}^{\mathrm{T}}\hat{\mathrm{V}}_{\mathrm{I}}^{-1}\mathrm{Z}})$, we can easily obtain the numeric values of $\mathrm{Var}(\hat{\gamma}_{\mathrm{EB_k}})$ through the equation $\mathrm{Var}(\hat{\gamma}_{\mathrm{EB_k}})$ = $\mathrm{Var}(\hat{\gamma}_{\mathrm{CSPML_k}})$+Var[f(Z)]+2$\mathrm{Cov}[\hat{\gamma}_{\mathrm{CSPML_k}}, \mathrm{f(Z)}]$, using the simulated values of Z and $\hat{\gamma}_{\mathrm{CSPML}}$. A similar idea can be applied to obtain $\mathrm{Cov}(\hat{\gamma}_{\mathrm{EB_j}}, \hat{\gamma}_{\mathrm{EB_k}})$. In the simulation results, we show that when $\hat{\gamma}_{\mathrm{CSPML}}$ differs from $\hat{\gamma}_{\mathrm{I}}$, i.e. $\mathrm{Z} \nrightarrow 0$, the impact on the variance calculation is moderately acceptable and lead to desirable results.

3.3 Simulation Studies

3.3.1 Simulation Settings

We evaluated the performance of our proposed estimators through simulation studies in various settings and compared it to MLE from the direct regression, CSPML estimators and individual EB estimators, which incorporate single external model information. In each simulation, we compared six methods (direct regression, CSPML, EB, IVW, OCWE and SC-Learner) considering both overall and covariate-wise performance. We used the estimated standard error (ESE) to assess the precision gain of point estimates compared with the direct regression, and evaluated three overall metrics on a validation dataset of size $\mathrm{N_{test}}$=1,000:

- Average estimated variance of logit-transformed predicted probability: $\bar{\mathrm{V}}[\mathrm{logit}(\hat{\mathrm{p}})] = \frac{1}{\mathrm{N_{test}}}\sum_{\mathrm{i}=1}^{\mathrm{N_{test}}}\hat{\mathrm{V}}\mathrm{ar}[\mathrm{logit}(\hat{\mathrm{p}}_{\mathrm{i}})]$, where $\hat{\mathrm{p}}$ denotes the estimated probability and $\mathrm{logit}(\mathrm{p}) = \log(\frac{\mathrm{p}}{1-\mathrm{p}}) = (\mathbf{X}, \mathrm{B})\boldsymbol{\gamma}$;
- Sum of squared errors: $\mathrm{SSE} = \frac{1}{\mathrm{N_{test}}}\sum_{\mathrm{i}=1}^{\mathrm{N_{test}}}(\hat{\mathrm{p}}_{\mathrm{i}} - \mathrm{p_{i0}})^2$, where $\hat{\mathrm{p}}_{\mathrm{i0}}$ denotes the true probability of $\mathrm{Y_i} = 1$ given $\mathrm{X_i}$ and $\mathrm{B_i}$;
- Scaled Brier score: $\mathrm{BS} = \frac{1}{\mathrm{N_{test}}}\sum_{\mathrm{i}=1}^{\mathrm{N_{test}}}(\mathrm{Y_i} - \hat{\mathrm{p}}_{\mathrm{i}})^2 / \frac{1}{\mathrm{N_{test}}}\sum_{\mathrm{i}=1}^{\mathrm{N_{test}}}(\mathrm{Y_i} - \bar{\mathrm{Y}})^2$.

A summary of all the simulation settings is listed in Figure 3.1. In the first four simulation scenarios (I, II, III, IV), we assumed that a logistic regression model with the following form could describe the relationship among a binary outcome Y and five covariates $(\mathrm{X_1}, \mathrm{X_2}, \mathrm{X_3}, \mathrm{X_4}, \mathrm{B})$, where $\mathrm{X_1}, \mathrm{X_2}, \mathrm{X_3}, \mathrm{X_4}$ had been used in at least one external model while B was only available in the current study: $\mathrm{logit}[\Pr(\mathrm{Y} = 1|\mathrm{X_1}, \mathrm{X_2}, \mathrm{X_3}, \mathrm{X_4}, \mathrm{B})] = -1 - 0.5\sum_{\mathrm{i}=1}^{4}\mathrm{X_i} + 0.5\mathrm{B}$. Here $\mathrm{X_1}, \mathrm{X_2}, \mathrm{X_3}, \mathrm{X_4}$ and B followed a standard multivariate normal distribution with 0.3 correlation and the prevalence of $\mathrm{Y} = 1$ was 0.32.

- Simulation I evaluated the idealized case where the internal and the external models were fitted on the homogeneous population.

- Simulation II assessed the performance of the proposed method when external model 1 had biased $\hat{\boldsymbol{\beta}}$ estimates fitted from $\mathrm{Y}|\mathrm{X}$, where we obtained the incorrect model estimates by fitting external model 1 on a small dataset of size 500.

- Simulation III aims to show the impact of heterogeneous covariate distribution (X, B) in the external population. As Estes et al. [2017] assessed, the disparity of the (X, B) distribution between the two populations can come from (i) a different conditional distribution $\mathrm{B}|\mathrm{X}$, (ii) a different marginal B distribution, (iii) a different marginal X distribution, or a combination of these reasons. In this simulation scenario, we showed the combination of (i) and (ii) as an example, but we will discuss the result of other scenarios in Simulation Results in Section 3.2.

- Simulation IV evaluated the situation where the outcome model was misspecified in external model 3.

We assumed that there was an internal study of size n=200 and three external models had been fitted to a very large synthetic dataset (sample size $\mathrm{m}_1 = \mathrm{m}_2 = \mathrm{m}_3 = 30,000$ for simplicity) that is sampled from the true data generating mechanism and gives precise estimates of the model parameters (except external model 1 in Simulation II). The external sample size need not be as large as 30,000 to achieve good performance as long as the estimated model parameters are close to the true parameters. Sensitivity analysis (results not shown) using external sample size $\mathrm{m}_1 = \mathrm{m}_2 = \mathrm{m}_3 = 1,000$ showed small numerical differences compared with m1= m2= m3=30,000.

In simulation V and VI, we considered the outcome model with higher dimension and homogeneous populations between the internal and the external models. In these two scenarios, the internal sample size n=500 was used.

- Simulation V evaluated the situation where the outcome model contained nine X's and one B: $\mathrm{logit}[\Pr(\mathrm{Y}=1|\mathrm{X}_1,...,\mathrm{X}_9,\mathrm{B})] = -1 - 0.5\sum_{\mathrm{i}=1}^{9}\mathrm{X}_\mathrm{i} + 0.5\mathrm{B}$. Specifically, external model 1 only contained two predictors, X_1 and X_2, external model 2 contained seven predictors, $\mathrm{X}_1,...,\mathrm{X}_7$, and external model 3 contained six predictors $\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3, \mathrm{X}_4, \mathrm{X}_7$ and X_8.

- simulation VI evaluated the situation where the full model contained three X's and five B's: $\mathrm{logit}[\Pr(\mathrm{Y}=1|\mathrm{X}_1,\mathrm{X}_2,\mathrm{X}_3,\mathrm{B}_1,...,\mathrm{B}_5)] = -1 - 0.5\sum_{\mathrm{i}=1}^{3}\mathrm{X}_\mathrm{i} + 0.5\sum_{\mathrm{i}=1}^{5}\mathrm{B}_\mathrm{i}$, using the same external models as simulation I.

		Internal Data	External Model 1	External Model 2	External Model 3
I: Correctly specified external models	Model	Y|X1, X2, X3, X4, B	Y|X1, X2	Y|X1, X3	Y|X1, X2, X3, X4
	Sample size	n=200	m_1=30,000	m_2=30,000	m_3=30,000
II: Biased estimate from external model 1	Model	Y|X1, X2, X3, X4, B	Y|X1, X2	Y|X1, X2	Y|X1, X2, X3, X4
	Sample size	n=200	m_1=500 (incorrect model estimates)	m_2=30,000	m_3=30,000
III: The joint distribution of (X, B) is different in the external model 1*	Model	Y|X1, X2, X3, X4, B	Y|X1, X2	Y|X1, X2	Y|X1, X2, X3, X4
	(X, B)	(X, B) ~N ((**0**, 0), 1), correlation 0.3	(X, B) ~ N ((**0**, 1.5), 1), correlation 0.8	Same as the internal data	Same as the internal data
	Sample size	n=200	m_1=30,000	m_2=30,000	m_3=30,000
IV: The outcome model is misspecified in the external model 3 †	Model	Y|X1, X2, X3, X4, B	Y|X1, X2	Y|X1, X3	Y|X1, X2, X3, X4
	Outcome model	Logit Pr(Y=1|X, B)= -1-0.5X+0.5B	Same as the internal data	Same as the internal data	Logit Pr(Y=1|X, B)= 1+0.5X+0.5B
	Sample size	n=200	m_1=30,000	m_2=30,000	m_3=30,000
V: homogeneous populations with correct external models (high dimensional X's)	Model	Y|X1, X2, X3, X4, X5, X6, X7, X8, X9, B	Y|X1, X2	Y|X1, X2, X3, X4, X5, X6	Y|X1, X2, X3, X4, X7, X8
	Sample size	n=500	m_1=30,000	m_2=30,000	m_3=30,000
VI: homogeneous populations with correct external models (high dimensional B's)	Model	Y|X1, X2, X3, B1, B2, B3, B4, B5	Y|X1, X2	Y|X1, X3	Y|X1, X2, X3, X4
	Sample size	n=500	m_1=30,000	m_2=30,000	m_3=30,000

* More scenarios that differ the joint distribution of (X, B) are presented in Appendix D of Supplementary Material.

† More scenarios that change the magnitude of outcome model are presented in Appendix D of Supplementary Material.

Note: If not specifically stated, (X, B) always followed Gaussian distribution with mean zero, standard deviation 1 and correlation 0.3; the outcome model used was the logistic model with the form logit Pr(Y=1|X, B)= -1-0.5X+0.5B.

Figure 3.1: Simulation settings snapshot

3.3.2 Simulation Results

In Table 3.1 simulation I, we see that all CSPML and EB estimators are unbiased as expected. The estimated standard error (ESE, in square brackets) accurately reflect the true standard deviation (SD, in round brackets) from 500 simulations. Both OCWE and SC-Learner had better overall performance than single EB estimators, while SC-Learner outperformed OCWE with respect to both the covariate-wise and the overall metrics (Figure 3.2).

Table 3.1: Results of simulation I–IV. Internal dataset had size n=200; green represents good performance with small bias and large efficiency gain, yellow represents underestimated ESE (in square brackets) compared with SD (in round brackets), and red represents poor performance of bias/95% coverage rate.

Simulation		Direct Regression	Internal Data + External 1		Internal Data + External 2		Internal Data + External 3		Composite of EB Estimators		
			CSPML 1	EB 1	CSPML 2	EB 2	CSPML 3	EB 3	IVW	OCWE	SC-Learner
		Bias (SD) [ESE] 95% Coverage Rate									
I	Weight	/	/	/	/	/	/	/	(.33, .33,.33)	(.27, .12, .61)	/
	γ_0	-.041 (.197) [.193] 95%	-.008 (.081) [.085] 96%	-.032 (.162) [.160] 95%	-.007 (.089) [.084] 95%	-.032 (.163) [.161] 95%	.004 (.050) [.085] 98%	-.033 (.169) [.168] 95%	-.032 (.164) [.162] 95%	-.031 (.166) [.164] 95%	-.032 (.164) [.162] 95%
	γ_1	-.020 (.229) [.208] 94%	-.004 (.095) [.097] 96%	-.014 (.190) [.173] 93%	-.002 (.089) [.097] 97%	-.013 (.188) [.174] 93%	-.002 (.066) [.074] 96%	-.016 (.198) [.181] 93%	-.014 (.191) [.176] 93%	-.016 (.194) [.178] 93%	-.014 (.191) [.175] 93%
	γ_2	-.017 (.219) [.208] 94%	-.008 (.091) [.096] 97%	-.014 (.180) [.173] 94%			-.003 (.062) [.210] 97%	-.014 (.188) [.182] 94%	-.015 (.195) [.187] 94%	-.011 (.190) [.181] 94%	-.014 (.184) [.176] 94%
	γ_3	-.018 (.208) [.207] 96%			-.002 (.090) [.095] 98%	-.013 (.171) [.174] 96%	-.002 (.062) [.174] 96%	-.015 (.177) [.182] 95%	-.015 (.185) [.185] 95%	-.012 (.184) [.186] 95%	-.014(.174) [.176] 95%
	γ_4	-.026 (.217) [.208] 94%					-.007 (.067) [.208] 97%	-.022 (.187) [.194] 96%	-.025 (.207) [.196] 94%	-.023 (.198) [.191] 95%	-.022 (.187) [.194] 96%
	γ_B	.016 (.219) [.207]									
II	Weight	/	/	/	/	/	/	/	[.33, .33, .34]	[.23, .28, .49]	/
	γ_0	-.042 (.198) [.193] 95%	.042 (.080) [.086] 87%	-.024 (.168) [.161] 94%	-.008 (.081) [.085] 96%	-.032 (.162) [.160] 95%	.003 (.050) [.058] 98%	-.034 (.170) [.169] 95%	-.031 (.166) [.162] 95%	-.031 (.167) [.163] 95%	-.031 (.166) [.162] 95%
	γ_1	-.020 (.229) [.208] 94%	.153 (.094) [.100] 65%	-.015 (.202) [.174] 92%	-.004 (.095) [.097] 96%	-.014 (.190) [.173] 93%	-.017 (.066) [.074] 97%	-.018 (.198) [.182] 94%	-.007 (.196) [.176] 93%	-.009 (.197) [.177] 92%	-.007 (.196) [.175] 93%
	γ_2	-.018 (.220) [.208] 94%	-.143 (.091) [.098] 70%	-.041 (.190) [.174] 93%	-.008 (.091) [.096] 96%	-.014 (.180) [.173] 94%	.009 (.062) [.073] 95%	-.013 (.188) [.183] 94%	-.022 (.185) [.176] 94%	-.020 (.186) [.177] 94%	-.028 (.189) [.177] 94%
	γ_3	-.019 (.208) [.207] 96%					.015 (.062) [.072] 94%	-.012 (.178) [.194] 97%	-.016 (.198) [.195] 95%	-.015 (.193) [.192] 95%	-.015 (.192) [.192] 95%
	γ_4	-.027 (.217) [.208] 94%					.004 (.067) [.073] 97%	-.021 (.187) [.195] 96%	-.025 (.207) [196] 94%	-.021 (.202) [.193] 94%	-.021 (.187) [.195] 96%
	γ_B	-.017 (.219) [.207] 98%									
III	Weight	/	/	/	/	/	/	/	[.33, .33, .33]	[.15, .35, .49]	/
	γ_0	-.041 (.198) [.193] 95%	.648 (.082) [.100] 0%	-.009 (.199) [.162] 91%	-.039 (.081) [.085] 97%	-.004 (.162) [.160] 95%	-.033 (.051) [.058] 98%	-.039 (.169) [.168] 95%	-.029 (.176) [.163] 93%	-.031 (.173) [.163] 93%	-.029 (.176) [.162] 93%
	γ_1	-.019 (.228) [.208] 94%	-.156 (.096) [.113] 75%	-.026 (.221) [.175] 89%	.012 (.095) [.097] 97%	-.010 (.189) [.173] 93%	.022 (.066) [.074] 97%	-.012 (.197) [.181] 94%	-.016 (.202) [.176] 91%	-.014 (.198) [.177] 93%	-.016 (.202) [.176] 91%
	γ_2	-.017 (.219) [.208] 95%	.160 (.091) [.113] 76%	-.024 (.211) [.175] 90%	-.015 (.091) [.096] 97%	-.015 (.179) [.173] 94%	-.005 (.062) [.073] 95%	-.014 (.188) [.182] 94%	-.017 (.192) [.176] 93%	-.016 (.188) [.177] 94%	-.019 (.200) [.177] 93%
	γ_3	-.019 (.209) [.207] 96%					-.014 (.062) [.072] 94%	-.014 (.178) [.182] 97%	-.018 (.198) [.195] 96%	-.017 (.193) [.192] 94%	-.018 (.192) [.192] 96%
	γ_4	-.026 (.217) [.208] 94%					-.021 (.067) [.073] 97%	-.021 (.188) [.194] 96%	-.026 (.207) [.196] 94%	-.025 (.202) [.193] 95%	-.025 (.188) [.195] 96%
	γ_B	.017 (.219) [.207] 98%									
IV	Weight	/	/	/	/	/	/	/	[.34, .33, .32]	[.57, .38,.04]	/
	γ_0	-.041 (.198) [.193] 95%	-.008 (.081) [.085] 96%	-.032 (.163) [.161] 94%	-.007 (.089) [.084] 95%	-.032 (.164) [.161] 95%	1.69 (.189) [.162] 0%	-.031 (.169) [.178] 96%	-.032 (.173) [.165] 93%	-.030 (.164) [.160] 93%	-.032 (.172) [.164] 93%
	γ_1	-.020 (.230) [.208] 94%	-.005 (.095) [.097] 96%	-.014 (.190) [.174] 93%	-.002 (.089) [.097] 97%	-.014 (.189) [.174] 93%	.747 (.178) [.177] 3%	-.016 (.197) [.192] 95%	-.015 (.201) [.178] 92%	-.013 (.191) [.173] 93%	-.015 (.200) [.178] 92%
	γ_2	-.018 (.220) [.208] 95%	-.008 (.091) [.096] 97%	-.014 (.180) [.175] 94%			.756 (.174) [.177] 3%	-.014 (.188) [.192] 94%	-.015 (.206) [.189] 93%	-.013 (.196) [.187] 94%	-.014 (.197) [.177] 93%
	γ_3	-.019 (.208) [.207] 96%			-.003 (.089) [.095] 98%	-.013 (.171) [.180] 96%	.762 (.167) [.175] 1%	-.014 (.178) [.191] 95%	-.015 (.194) [.187] 95%	-.013 (.193) [.189] 94%	-.014 (.187) [.181] 95%
	γ_4	-.026 (.216) [.208] 94%					.746 (.179) [.176] 4%	-.022 (.188) [.193] 93%	-.025 (.216) [.202] 94%	-.026 (.217) [.207] 95%	-.022 (.216) [.193] 94%
	γ_B	.019 (.218) [.207] 98%									

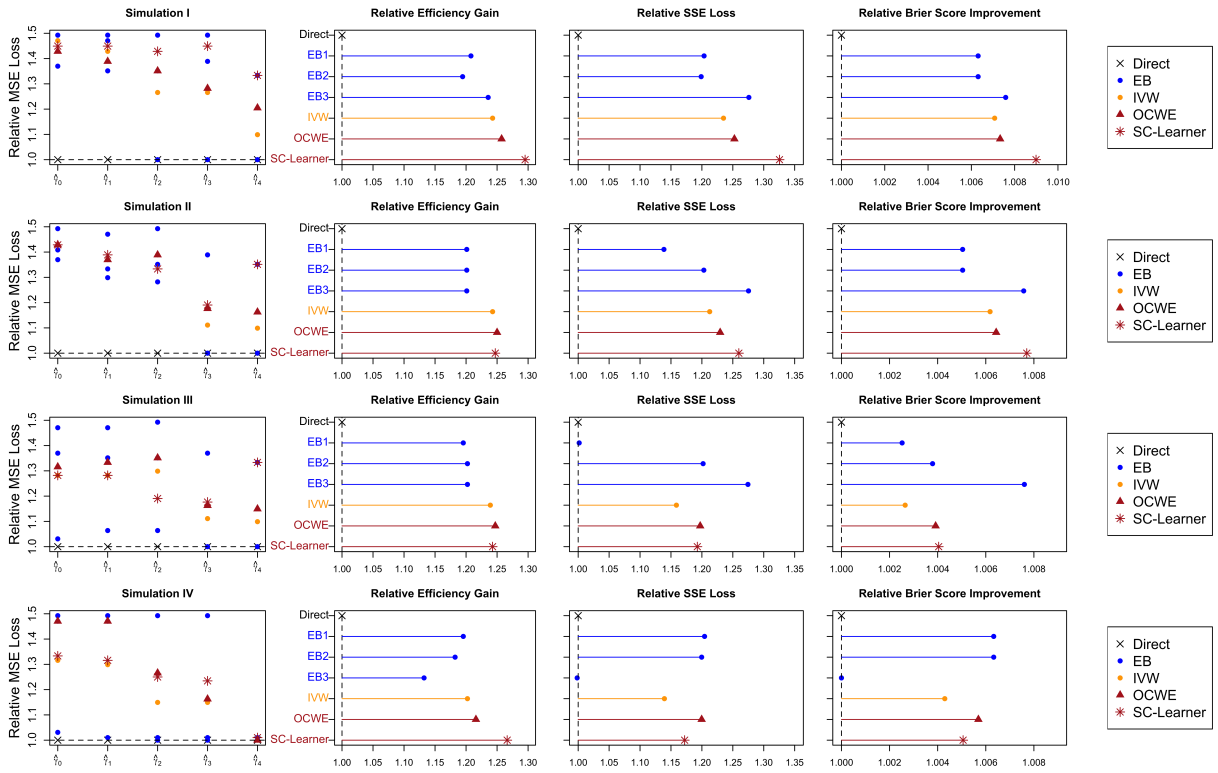


Figure 3.2: Visualization of the metrics to evaluate the performance in simulations I-IV. Scatter plot shows the covariate-wise relative MSE improvement compared with the direct regression (x axis represents $\hat{\gamma}_{\mathrm{X}}$; $\hat{\gamma}_{\mathrm{B}}$ not shown since no external information incorporated; larger y values represent larger MSE loss) while the line plots represent the relative efficiency/SSE/BS improvement compared with the direct regression fitting on a validation dataset of size 1,000 (longer lines represent larger improvement).

In Table 3.1 simulation II, CSPML_1 was biased with poor 95% coverage rate due to biased estimation of $\hat{\boldsymbol{\beta}}$. Although trading off most of the precision gain, EB_1 corrected the bias in CSPML_1. The bias of CSPML_1 also caused underestimation of the standard error of EB_1 as highlighted in yellow, which was because $\hat{\boldsymbol{\gamma}}_{\mathrm{CSPML}_1}$ differed from $\hat{\boldsymbol{\gamma}}_{\mathrm{I}}$ as pointed out in Section 3.2.4. Despite that, both OCWE and SC-Learner detected the least efficiency gain of EB_3 among all EB estimators and provided unbiased overall estimates. Moreover, OCWE and SC-Learner had the largest relative efficiency gain compared with the direct regression than other methods, as well as the best performance regarding the relative SSE and BS gain only second to EB_3 (Figure 3.2).

Results of simulation III and IV in Table 3.1 indicates that when incorporating discordant external information from heterogeneous populations (one with different distribution of the joint [X,B] and another with different outcome model), EB estimators are able to correct the bias of CSPML estimators by trading off some precision gain, and the proposed estimators could further identify and down-weight that particular EB estimators. Estes et al. [2017] provided comprehensive sim-

ulation results to show that the EB estimator can protect against the bias due to heterogeneous (X, B) distribution from the external model. Their results indicated that CSPML estimator would have a substantial bias when the difference came from the conditional distribution $\mathrm{B}|\mathrm{X}$ or marginal distribution B, and marginal distribution X when the X-B interaction term is involved in the true $\mathrm{Y}|\mathrm{X}$, B model. When evaluating these scenarios (summary and results in Appendix B.1), we found a similar pattern and conclusion as simulation III. In summary, these simulation results provided evidence that the proposed approach is robust to the heterogeneous covariate distribution (X, B) of the external population.

In addition to simulation IV, we have assessed different forms of misspecified outcome model in the external population listed in Appendix B.4, including different intercept only and different X coefficients only, which showed similar conclusions. Simulation IV also showed that although the higher dimension of the external model will lead to a better overall prediction when the internal and external has the same population (simulation I and II), it is not the case when the transportability assumption is violated: OCWE identified that external model 3 came from a different distribution than the internal study and thus assigned the smallest weight to EB_3, even though it had the greatest number of predictors compared with the other two external models. In both simulations, OCWE and SC-Learner had similar covariate-wise and overall performances (Figure 3.2).

Table 3.2 further shows that the proposed estimators have decent performance when the number of predictors in the external models has large differences (simulation V) and when the dimension of B is larger than X's (simulation VI). In simulation V, compared with other external models that included more than six predictors, external model 1 only used two predictors and thus provided the least amount of extra information. We see that OCWE assigns the minimum weight to EB_1 and achieves the largest relative efficiency gain among all estimators (Figure 3.3). The fact that SC-Learner outperformed OCWE with respect to decreasing the covariate-wise variance of $\hat{\gamma}_5$ to $\hat{\gamma}_9$ reveals that SC-Learner is flexible enough to select the external information on the covariate-level when $\mathrm{p_k}$ (the number of predictors used in the $\mathrm{k^{th}}$ external model) is very different from others. Simulation VI showed that when the dimension of B was much larger than X, the overall benefit of combining multiple EB estimators would be limited due to the small amount of external information added from the external models.

Table 3.2: Results of simulation V–VI. Internal dataset had size n=500; green represents good performance with small bias and large efficiency gain.

Simulation		Direct Regression	Internal Data + External 1		Internal Data + External 2		Internal Data + External 3		Composite of EB Estimators		
			CSPML 1	EB 1	CSPML 2	EB 2	CSPML 3	EB 3	IVW	OCWE	SC-Learner
		Bias (SD) [ESE] 95% Coverage Rate									
V	Weight	/	/	/	/	/	/	/	(.33, .33,.33)	(.13, .432, .429)	/
	γ_0	-.035 (.153) [.149] 94%	.02 (.100) [.098] 95%	.031 (.132) [.129] 95%	-.002 (.054) [.059] 97%	-.031 (.139) [.137] 94%	-.010 (.064) [.066] 96%	-.031 (.137) [.135] 94%	-.031 (.135) [.132] 94%	-.031 (.137) [.133] 94%	-.031 (.135) [.132] 94%
	γ_1	-.023 (.160) [.158] 94%	.002 (.089) [.094] 96%	.016 (.133) [.133] 94%	.010 (.055) [.061] 97%	-.019 (.144) [.144] 94%	.009 (.061) [.067] 97%	-.019 (.142) [.142] 94%	-.018 (.139) [.139] 94%	-.019 (.141) [.141] 94%	-.018 (.139) [.139] 94%
	γ_2	-.027 (.153) [.158] 95%	.014 (.094) [.094] 95%	.023 (.128) [.133] 95%	-.006 (.056) [.060] 97%	-.024 (.137) [.144] 94%	.013 (.064) [.066] 96%	-.025 (.136) [.142] 95%	-.024 (.133) [.139] 95%	-.024 (.135) [.141] 95%	-.024 (.133) [.138] 95%
	γ_3	-.037 (.161) [.158] 95%			.002 (.057) [.061] 97%	-.031 (.144) [.145] 95%	.002 (.064) [.067] 97%	-.031 (.143) [.143] 95%	-.033 (.149) [.148] 95%	-.031 (.145) [.145] 95%	-.031 (.143) [.143] 95%
	γ_4	-.014 (.162) [.157] 94%			.025 (.054) [.060] 97%	-.014 (.145) [.144] 95%	.025 (.062) [.066] 96%	-.016 (.144) [.142] 95%	-.015 (.150) [.147] 95%	-.014 (.147) [.144] 95%	-.015 (.145) [.142] 96%
	γ_5	-.014 (.160) [.158] 96%			.010 (.055) [.060] 97%	-.011 (.143) [.153] 95%			-.013 (.154) [.150] 95%	-.012 (.153) [.150] 95%	-.011 (.143) [.153] 96%
	γ_6	-.023 (.160) [.158] 95%			-.010 (.057) [.061] 99%	-.021 (.143) [.153] 96%			-.023 (.154) [.151] 95%	-.022 (.152) [.151] 95%	-.021 (.143) [.153] 96%
	γ_7	-.015 (.165) [.158] 95%			-.008 (.056) [.060] 99%	-.014 (.148) [.153] 96%			-.015 (.159) [.150] 95%	-.015 (.157) [.150] 95%	-.014 (.148) [.153] 96%
	γ_8	-.033 (.159) [.159] 95%					-.001 (.063) [.067] 98%	-.029 (.141) [.151] 96%	-.031 (.153) [.152] 97%	-.031 (.151) [.151] 95%	-.029 (.141) [.151] 97%
	γ_9	-.027 (.154) [.154] 97%					-.021 (.066) [.067] 97%	-.026 (.136) [.151] 97%	-.026 (.148) [.151] 97%	-.026 (.146) [.150] 95%	-.026 (.136) [.151] 97%
	γ_B	.034 (.172) [.172] 97%									
VI	Weight	/	/	/	/	/	/	/	[.33, .33, .34]	[.01, .0004, .987]	/
	γ_0	-.027 (.126) [.121] 94%	.003 (.078) [.073] 92%	-.020 (.108) [.103] 94%	.003 (.078) [.073] 93%	-.031 (.108) [.103] 94%	.005 (.078) [.073] 92%	-.020 (.110) [.105] 94%	-.020 (.109) [.104] 94%	-.020 (.110) [.105] 94%	-.020 (.109) [.104] 94%
	γ_1	-.012 (.132) [.132] 95%	-.015 (.088) [.085] 95%	-.013 (.115) [.115] 95%	-.019 (.086) [.085] 94%	-.019 (.115) [.115] 95%	-.019 (.084) [.081] 95%	-.013 (.116) [.116] 95%	-.013 (.115) [.115] 95%	-.013 (.116) [.116] 95%	-.013 (.115) [.115] 95%
	γ_2	-.004 (.137) [.132] 93%	.005 (.087) [.085] 95%	.004 (.118) [.115] 93%			.004 (.084) [.081] 94%	.003 (.119) [.117] 94%	.004 (.124) [.121] 93%	.003 (.119) [.117] 94%	.004 (.118) [.116] 94%
	γ_3	-.011 (.131) [.132] 96%			.005 (.085) [.085] 95%	-.007 (.113) [.117] 96%	.005 (.081) [.082] 94%	-.007 (.114) [.118] 96%	-.008 (.119) [.121] 96%	-.007 (.114) [.118] 96%	-.007 (.113) [.117] 96%
	γ_{B1}	.007 (.126) [.121] 95%									
	γ_{B2}	-.025 (.137) [.132] 96%									
	γ_{B3}	-.014 (.135) [.132] 97%									
	γ_{B4}	-.012 (.131) [.132] 96%									
	γ_{B5}	-.013 (.139) [.132] 95%									

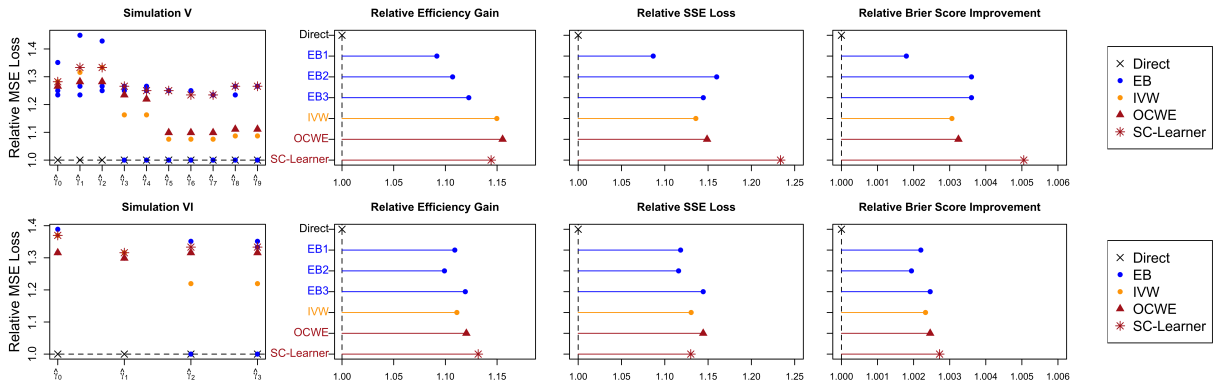


Figure 3.3: Visualization of the metrics to evaluate the performance in simulations V-VI. Scatter plot shows the covariate-wise relative MSE improvement compared with the direct regression (x axis represents $\hat{\gamma}_{\mathrm{X}}$; $\hat{\gamma}_{\mathrm{B}}$ not shown since no external information incorporated; larger y values represent larger MSE loss) while the line plots represent the relative efficiency/SSE/BS improvement compared with the direct regression fitting on a validation dataset of size 1,000 (longer lines represent larger improvement).

3.4 Application to Prostate Cancer Data

To assess the performance of the proposed estimators in a real data example, we developed a model for predicting the risk of high-grade prostate cancer (Gleason score > 6) using a combination of a set of internal individual-level data and two external risk calculators from different studies. The first risk calculator was developed based on the Prostate Cancer Prevention Trial in the United States [Thompson et al., 2006]. This calculator, denoted as PCPThg, is built on five clinical variables including prostate-specific antigen (PSA) level, digital rectal examination (DRE) findings, age, race (African American or not) and prior biopsy results using the following model:

$$\mathrm{logit}(\mathrm{p_i}) = -6.25 + 1.29\mathrm{log}(\mathrm{PSA_i}) + \mathrm{DRE_i} + 0.03\mathrm{Age_i} + 0.96\mathrm{Race_i} - 0.36\mathrm{Biopsy_i}, \quad (3.3)$$

where $\mathrm{p_i}$ is the probability of observing high-grade prostate cancer for subject i. The second risk calculator is the European Randomized Study of Screening for Prostate Cancer (ERSPC) risk calculator 3 [Roobol et al., 2012], which uses slightly different clinical variables to predict the same risk as PCPThg: PSA level, DRE findings, and transrectal ultrasound prostate volume (TRUS-PV) in a logistic regression model shown as below:

$$\mathrm{logit}(\mathrm{p_i}) = -3.15 + 1.18\mathrm{log_2}(\mathrm{PSA_i}) + 1.81\mathrm{DRE_i} - 1.51\mathrm{log_2}(\text{TRUS-PV}_\mathrm{i}), \quad (3.4)$$

where TRUS-PV was categorized as a 3-category variable described in Roobol et al. [2012]. In addition to all the predictors used in the external model 3.3 and 3.4, we considered adding two more log-2-transformed biomarkers that had not been widely used but shown to be predictive of prostate cancer [Tomlins et al., 2015, Truong et al., 2013], prostate cancer antigen 3 (PCA3) and TMPRSS2:ERG (T2:ERG) gene fusions to our target model:

$$\begin{aligned}\text{logit}(\text{p}_\text{i}) = \gamma_0 &+ \gamma_1 \log_2(\text{PSA}_\text{i}) + \gamma_2 \text{DRE}_\text{i} + \gamma_3 \text{Age}_\text{i} + \gamma_4 \text{Biopsy}_\text{i} + \gamma_5 \text{Race}_\text{i} \\ &+ \gamma_6 \log_2(\textbf{TRUS-PV}_\text{i}) + \gamma_7 \log_2(\text{PCA3}_\text{i} + 1) + \gamma_8 \log_2(\textbf{T2:ERG}_\text{i} + 1), \end{aligned} \tag{3.5}$$

Using the data from Tomlins et al. [2015], we had 678 male patients in the internal dataset who had complete data of all eight covariates listed in model 3.5 and an independent validation data of size 1,174 (sample size reduced from the initial 679 and 1218 patients, respectively due to missing TRUS-PV that was not previously used in Tomlins et al. [2015]). Details of the individual-level data, including the description of the internal and the validation dataset, and the recent applications using the same setting can be found in Tomlins et al. [2015] and Cheng et al. [2019].

PCPThg utilizes standard clinical and demographic variables that have been widely used while ERSPC additionally considers the prostate volume that was shown to be related to PSA level [Bohnen et al., 2007]. In addition, similarities of prostate-specific antigen patterns between the United States and European populations prostate-specific antigen patterns have been shown [Simpkin et al., 2016]. Therefore, incorporating information from both risk calculators can potentially provide more accurate estimation of the risk parameters and narrower confidence bands, which could in turn yield better prediction performance and improved inference.

In order to reconcile the discrepancy between the covariates used in the external models and have a compatible interpretation of the intercept, we centered all variables in the original models 3.3 and 3.4, and log2-transformed PSA and TRUS-PV by adjusting the corresponding intercepts a-priori (details in Appendix B.5). We present the estimated coefficients and standard errors in Table 3.3. Similar to the simulation study, we calculated the scaled Brier Score and the average estimated variance of logit-transformed predicted probability based on the validation dataset as the prediction metrics.

As shown in Table 3.3, OCWE assigns almost zero weight to the ERSPC, which indicated a large population discrepancy between the internal data and the underlying European population possibly due to the difference in the intercept, which reflects that the prevalence of high-grade prostate cancer is higher in the European population compared with patients in the United States who had average covariate values. Even though, SC-Learner was able to make the most of the little improvement provided by ERSPC and augment the point-wise precision gain for covariate PSA, DRE and TRUS-PV (3%, 4% and 12% more compared with OCWE, respectively), which

led to the largest overall improvement as well (17.2 % decrease of the average prediction variance compared with the direct regression).

3.5 Discussion

The proposed framework along with two weighted estimators, OCWE and SC-Learner, adds to the evolving research on using external summary-level information to bolster the statistical efficiency of the internal study for improved inference. This new method is flexible and robust in the ways that (i) it is capable of incorporating external models that use a slightly different set of covariates; (ii) it is able to identify the most relevant external information and diminish the influence of information that is less compatible with the internal data; and (iii) it nicely balances the bias-variance trade off while preserving the most precision gain. Moreover, our extensive simulation studies and the real data example show that the proposed estimators are more efficient and robust than the naïve analysis of the internal data and other naïve combinations of external estimators in both idealized and non-idealized settings.

Compared with a single EB estimator, the proposed composite estimators can have up to 32% more improvement in MSE regarding one covariate (Figure 3.2) and decent improvement regarding the overall metric such as 20% further decrease in SSE and 11.5% further decrease in estimated prediction variance (Figure 3.2). In some cases, several single EB estimators that showed limited gains mitigate another single EB estimator's excellent performance during integration, e.g. Simulation II. In reality, the proposed composite estimators will be preferred over a single EB estimator, since it is often difficult to pick the best external model that contains the most useful information to boost the inference of the internal study among several available external models.

In practice, the choice of SC-Learner or OCWE mainly depends on the features of the external models and the user's research goal. As shown in simulation V, if at least one of the external models used very few predictors compared to the full dimension of (X, B), i.e., $p_k << p$, we suggest using SC-Learner as it can adapt the external information being considered covariate-wise and thus prevent the gain in certain covariates from being washed away when the dimension of predictors are uneven. Similarly, we would recommend using SC-Learner if the researcher cares about maximizing the precision gain on the covariate-level or improving precision in certain covariates are of particular interest. On the other hand, if the research goal involves ranking the usefulness/relevancy of the external model, OCWE would be a good choice as it provides one unified weight for all covariates in the same model and outperforms comparable estimators such as IVW (inverse variance-weighed estimator).

A different approach to estimating the weights in OCWE would be cross-validation. For example, instead of minimizing the estimated prediction variance over all internal observations, one

Table 3.3: Results of the real data example for predicting the risk of high-grade prostate cancer. Internal dataset of size n=678; validation dataset of size N_{test}=1,174; $\downarrow$ %, percentage of ESE decrease compared with the direct regression; Firth correction applied in direct regression, and estimated PCPThg and ERSPC; REF, reference; green represents good performance with small bias and large efficiency gain, and grey represents no efficiency improvement due to no added information from the external calculator.

	Direct Regression	PCPThg		ERSPC		Internal data + PCPThg		Internal data + ERSPC		Composite of EB Estimators		
		Original	Estimated	Original	Estimated	CSPML 1	EB 1	CSPML 2	EB 2	IVW	OCWE	SC-Learner
Weight	/	/	/	/	/	/	/	/	/	[.5, .5]	[.99, .00]	/
$\bar{V}[\text{logit}(\hat{p})]$	.1199	/	.065	/	.048	.073	.104	.078	.105	.104 ($\downarrow$ 13.2%)	.104 ($\downarrow$ 13.2%)	.0996 ($\downarrow$ 17.2%)
Brier Score	.902	1.059	.987	.954	.994	.942	.901	.847	.901	.901 ($\downarrow$ .12%)	.901 ($\downarrow$ .12%)	.901 ($\downarrow$ .12%)
	Point estimate (ESE) $\downarrow$% ESE w.r.t direct regression											
Intercept	-4.123 (.449)	-3.686268	-1.395 (.115)	-3.16	-1.382 (.111)	-6.422 (.437)	-4.128 (.445)	-5.989 (.437)	-4.130 (.444)	-4.129 (.443)	-4.128 (.445)	-4.129 (.443)
	REF	/	74%	/	75%	3%	1%	3%	1%	1%	1%	1%
$\log_2$(PSA)	0.860 (.144)	0.8941599	0.721 (.123)	1.175573	.878 (.123)	0.893 (.103)	0.860 (.132)	1.159 (.081)	0.861 (.124)	0.860 (.127)	0.860 (.130)	0.860 (.127)
	REF	/	14%	/	14%	28%	8%	44%	14%	11%	9%	12%
DRE	1.028 (.297)	1	1.134 (.256)	1.813195	1.298 (.269)	0.687 (.218)	1.027 (.269)	1.430 (.170)	1.029 (.253)	1.028 (.259)	1.027 (.269)	1.028 (.258)
	REF	/	14%	/	9%	26%	9%	43%	15%	13%	9%	13%
Age	0.0315 (.014)	0.03	0.0328 (.012)			0.0304 (.009)	0.0315 (.013)	0.032 (.014)	0.031 (.014)	0.032 (.014)	0.032 (.013)	0.032 (.013)
	REF	/	14%			31%	8%			1%	8%	8%
Biopsy	-1.165 (.290)	-0.36	-1.413 (.270)			-0.0286 (.158)	-1.163 (.255)	-1.187 (.290)	-1.165 (.291)	-1.164 (.272)	-1.163 (.255)	-1.163 (.255)
	REF	/	7%			46%	12%			6%	12%	12%
Race	0.193 (.329)	0.96	0.448 (.287)			0.819 (.218)	0.194 (.291)	0.160 (.329)	0.193 (.329)	0.193 (.308)	0.194 (.291)	0.194 (.291)
	REF	/	13%			34%	12%			6%	12%	12%
$\log_2$ (TRUS-PV)	-1.663 (.252)			-1.514128	-1.681 (.224)	-1.683 (.252)	-1.663 (.252)	-1.491 (.155)	-1.663 (.223)	-1.663 (.236)	-1.663 (.252)	-1.663 (.223)
	REF			/	11%			38%	12%	6%	0%	12%
$\log_2$ (PCA3+1)	.485 (.088)					0.495 (.088)	0.485 (.088)	0.507 (.088)	0.485 (.088)	0.485 (.088)	0.485 (.088)	0.486 (.088)
	REF											
$\log_2$ (T2:ERG+1)	.096 (.037)					0.095 (.037)	0.096 (.037)	0.099 (.037)	0.096 (.037)	0.096 (.037)	0.096 (.037)	0.096 (.037)
	REF											

could randomly split the data into training and testing data, and choose the weights that minimize the objective function over the testing data only, then average this over different data splits. This approach could potentially prevent overfitting, give more stable estimates of the weights and improve the predictive performance.

As is typical of shrinkage estimators, in finite-sized samples, the EB method sacrifices a small amount of efficiency when the assumption of full transportability is satisfied, but reduces the potential bias of the CSPML when full transportability is not satisfied, while still being more efficient than the simple estimate from the internal data. In the proposed method, the magnitude of the precision gain depends on the degree of the distributional similarity between the internal and external populations, i.e., the more similar, the more benefit we will gain by incorporating the external models. In the extreme case when these populations are completely different, our approach is very similar to analyzing the internal data only. On the contrary, when these populations share the identical joint distribution of (Y, X, B), we will achieve near to the maximum possible benefit.

Note that the proposed framework is not suitable if some predictors used in the external models are completely unmeasured in the internal study. In addition, the proposed framework is constructed based on parametric regression models, which requires the exact form of the external models and common covariates shared across different external models. In some cases, such as the real data example in this study, the authors were able to reconcile the discrepant transformations (i.e. one used natural-log PSA and the other used mean-centered log-2 based PSA) by reparametrizing the intercept. But this may not always be feasible, in which case we suggest considering methods that only require the predicted probability or the ability to estimate the probability given predictors without knowing the exact models, such as the synthetic data method proposed in Chapter 2. It is plausible that the assumed parametric model is not a good approximation of the internal data's underlying distribution. In the simulation (results not shown here), we saw that when the effect of non-linear terms was small, the proposed method could still correctly estimate the main effect. Besides, some reassurance about the selected structure of the parametric internal model can be obtained from the external models. The external models determine the X variables that are included and how they are included. For example, in our prostate cancer example, the external models took a log transformation of PSA, thus the parametric model for the internal data also includes a log transformation of PSA. Since the external datasets are typically large, we might surmise that if a large non-linearity or a strong interaction amongst the X variables were needed, it would have been included in the external model.

Recently, Zhang et al. [2020] proposed a general framework as the extension of Chatterjee et al. [2016] to solve the same genre of problem when the external parameter uncertainty cannot be ignored. In the situation where the external study population differs from the internal one, the performance of their method depends on the availability of high-quality reference data from the

external population, similar to in Chatterjee et al. [2016]. Our method also provides the option of incorporating external parameter uncertainty but more importantly it provides valid internal inference when it is in general hard to obtain the right reference data in reality. In addition, Kundu et al. [2019] proposed a generalized meta-analysis framework building from the Chatterjee et al. [2016] approach to combine information of multiple regression models with disparate covariates using a method of moment approach. Different from our goal, their method is an extension of the fixed-effect meta-analysis that also relies on the existence of reference data, and the performance of the proposed estimator depends on the quality and representativeness of the reference data.

One possible extension of the proposed method is the application in causal inference. If the treatment indicator was available as one of the X covariates, one could directly calculate the estimated average causal effect through the formula $\mathrm{E}_{\mathrm{X,B}}[\mathrm{E}(\mathrm{Y}|\mathrm{treatment} = 1, \mathrm{X}, \mathrm{B}) - \mathrm{E}(\mathrm{Y}|\mathrm{treatment} = 0, \mathrm{X}, \mathrm{B})]$ using the regression estimates obtained from the proposed method. Yang and Ding [2020a] considered a similar setting, where they view the internal data as the validation data with richer covariates while the external data serves as the main dataset with fewer covariates, aiming to improve the efficiency of the initial estimator $\hat{\boldsymbol{\gamma}}_{\mathrm{I}}$ from the internal dataset by incorporating a constructed zero-mean error-prone estimator $\hat{\boldsymbol{\beta}}_{\mathrm{I}} - \hat{\boldsymbol{\beta}}_{\mathrm{E}}$, where $\hat{\boldsymbol{\beta}}_{\mathrm{I}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{E}}$ are the estimators using X only from the internal and external population, respectively. Using notation from Estes et al. [2017], we can also reparametrize Yang and Ding's estimator as a weighted average of $\hat{\boldsymbol{\gamma}}_{\mathrm{I}}$ and $\hat{\boldsymbol{\gamma}}_{\mathrm{CSPML}}$, where the only difference is that the EB estimator has the shrinkage effect by empirically estimating the variance-covariance matrix that plays an important role in the weights.

It is worth noting that there is a popular field in machine learning called ensemble learning with a large and evolving literature, aiming to combine several base models to produce the optimal predictive model. Some representative ensemble methods include but not limited to Boosting [Schapire, 1990], Bagging [Freund and Schapire, 1997] and Stacking [Breiman, 1996] with some examples being random forest [Breiman, 2001] and Super Learner [van der Laan et al., 2007]. The key difference of our proposed method is that we have a specific parametric model of interest, and we are taking the weighted average of the estimated coefficients of that model from several estimators such that we can measure the impact of each predictor and its uncertainty, instead of directly weighting the predicted outcomes as in these ensemble methods. The proposed method can provide competitive and robust estimators for statistical inference. Moreover, the proposed estimators have improved efficiency compared with direct regression using the internal data only or naïve inverse variance-weighted estimator. However, if the research goal is to find the optimal predictive model with the minimum prediction error solely, especially when the underlying mechanism is not of interest, it would be worthwhile to explore the field of ensemble methods mentioned above, which is beyond the discussion of this study.

Last but not least, the issue of transportability of risk prediction models is a critical one and

one that is often encountered in practice. While it is plausible that the association between pairs of variables or even the joint distribution of all the variables is similar between populations, it is also plausible that they could differ, not just due to biological or behavioral differences in populations but also due to being collected in different parts of the world or different decades. The EB strategy will be a good choice, balancing between bias and efficiency when one is unsure about whether transportability assumptions hold for risk models across time, space or cohorts.

3.6 Software and Publication

R package MetaIntegration is available at https://github.com/umich-biostatistics/MetaIntegration.
The content of this chapter has been published in *Biostatistics* at
DOI: 10.1093/biostatistics/kxab017.

CHAPTER 4

Regression Inference for Multiple Populations by Integrating Summary-Level Data using Stacked Imputations

4.1 Introduction

In Chapter 3, we proposed a meta-inference framework using an empirical Bayes approach, which adaptively weighted the estimates from multiple external models to incorporate the most compatible information with the internal data while balanced the bias-variance trade-off. In this chapter, we revisit and extend the synthetic data method proposed in Chapter 2, and further allow for heterogeneity of covariate effects across the external populations, one of the key challenges in data integration and ignoring which would lead to potential estimation bias and misleading inference.

Efforts have been made to address the issue of population difference across studies. The meta-inference framework proposed in Chapter 3 assigns larger weights to the more compatible external data sources to incorporate valid supplementary information into the internal study. Chen et al. [2020] used a penalty function to identify the difference of aggregate information among data sources. Yang and Ding [2020b] employed a sensitivity parameter to quantify such systematic differences. However, the facts that different external studies may use different subsets of covariates and the underlying prediction model may be parametric or constructed by machine learning approaches add to the difficulties when making valid inference of both the internal and the external populations in data integration, let alone the external summary-level information may contain estimated regression coefficients or fitted predictions.

Although some of the existing approaches have considered heterogeneity across data sources, the main focus has been on improving the statistical efficiency of the internal dataset with little attempts to make statistical inference on external populations or allowing heterogeneous covariate effects across data sources. Wang, Wang, and Song [2012a] proposed a joint estimating procedure to merge longitudinal datasets while allowing different study-specific coefficients. The meta-

analysis approach proposed by Kundu, Tang, and Chatterjee [2019] used a generalized method of moments to estimate study-specific effects but only allows covariates that were measured in at least one of the external studies. Antonelli, Zigler, and Dominici [2017] proposed a unified Bayesian imputation framework built upon the work by Wang, Parmigiani, and Dominici [2012b], taking into account the prior odds of including a predictor in the outcome model given that it is in the exposure model, and allowing heterogeneous treatment effects by positing different population indicators in the outcome model.

In this chapter, we consider the situation where moderately sized individual data is available from the internal study, and there are K populations ($K \geq 1$), each of which provides some information about the relationship between the same outcome and a slightly different set of predictors. We propose an imputation-based methodology where the goal is to fit an outcome regression model with all available variables in the internal study while utilizing summary information from external models that may have used only a subset of the predictors. The method allows for heterogeneity of covariate effects across the external populations, by first generating synthetic outcome data in each population, then using stacked multiple imputation to create a long dataset with complete covariate information, and finally analyzing the imputed data with weighted regression. This flexible and unified approach attains the following four objectives: (i) incorporating supplementary information from a broad class of externally fitted predictive models or established risk calculators based on parametric regression or machine learning methods, as long as the external model can generate outcome values given covariates; (ii) improving statistical efficiency of the estimated coefficients in the internal study; (iii) improving predictions by utilizing even partial information available from prediction models that uses a subset of the full set of covariates used in the internal study; and (iv) providing valid statistical inference for the external population with potentially different covariate effects from the internal population.

The rest of this chapter is organized as follows: In Section 4.2, we introduce the proposed methodology. In Section 4.3, we evaluate the performance of our proposed approach in a simulation study. In Section 4.4, we apply the proposed strategy to a data example, where we build an expanded risk model to predict high-grade prostate cancer borrowing information from two existing risk prediction models. Concluding remarks are presented in Section 4.5.

4.2 Models and Methods

4.2.1 Notation

Let Y denote the outcome variable of interest, either continuous or binary. Consider $\mathbf{X}$ a set of P routinely measured variables and $\mathbf{B}$ a set of Q new variables, e.g. newly discovered biomarkers,

where $\mathbf{B}$ is only available in the internal study (i.e. $\mathbf{B}$ are unmeasured variables in all of the external studies). Let $\mathrm{S_k}$, k=0,...,K, denote the indicators of the K+1 study populations, where $\mathrm{S_0}$ represents the internal population and $\mathrm{S_k}$'s represents the K external populations.

We assume that a moderate dataset of size n with complete variables $\mathrm{Y}, \mathbf{X}$ and $\mathbf{B}$ is available to us from the internal study. For each external population $\mathrm{k} \geq 1$, a well-established reduced model for the same outcome Y is also available, each of which may use a slightly different set of predictors $\mathbf{X_k}$, a subset of $\mathbf{X}$. For example, if linear regression is used, the prediction model may look like: $\mathrm{E}(\mathrm{Y}|\mathbf{X_k}) = \beta_0 + \mathbf{X_k}\boldsymbol{\beta_k}$, where the dimension of $\mathbf{X_k}$ is $\mathrm{P_k} \leq \mathrm{P}$. We do not have access to the underlying individual-level data that was used to fit the external model but only the summary information. This summary information can come in different forms that we summarize into two categories:

Category 1: Directly available in the form of an externally fitted parametric regression model, along with the estimated model parameters $\hat{\boldsymbol{\beta}}_\mathrm{k}$;

Category 2: Any parametric or non-parametric models without knowing the exact form e.g. established risk calculators that provide the risk probability $\mathrm{P}(\mathrm{Y} = 1|\mathbf{X_k})$, for any $\mathbf{X_k}$.

We assume the target model of interest $\mathrm{Y}|\mathbf{X}, \mathbf{B}, \mathbf{S}$ is a generalized linear model (GLM):

$$g[\mathrm{E}(\mathrm{Y}|\mathbf{X}, \mathbf{B}, \mathbf{S})] = \gamma_0^{\mathrm{S_0}} + \sum_{\mathrm{k}=1}^{\mathrm{K}} \gamma_0^{\mathrm{S_k}} \mathrm{S_k} + \sum_{\mathrm{p}=1}^{\mathrm{P}} \gamma_{\mathrm{X_p}}^{\mathrm{S_0}} \mathrm{X_p} + \sum_{\mathrm{k}=1}^{\mathrm{K}} \sum_{\mathrm{p}=1}^{\mathrm{P}} \gamma_{\mathrm{X_p}}^{\mathrm{S_k}} \mathrm{S_k} \mathrm{X_p} + \sum_{\mathrm{q}=1}^{\mathrm{Q}} \gamma_{\mathrm{B_q}}^{\mathrm{S_0}} \mathrm{B_q}, \qquad (4.1)$$

where g is a known link function and the terms that contain $\mathrm{S_k}$ indicate the changes from the internal study. We assume that the distribution of $\mathrm{Y}|\mathbf{X}, \mathbf{B}, \mathbf{S}$ is correctly specified, which indicates that each external population can potentially differ in intercept and $\mathbf{X}$ covariate effect as long as those $\mathbf{X}$'s were used in the external model. For covariates that are unmeasured in the $\mathrm{k^{th}}$ external model (partial $\mathbf{X}$ and $\mathbf{B}$), we assume the covariate effects are the same as the internal study, i.e. no such population-interaction terms in model 4.1. In practice, we would force some $\gamma_{\mathrm{X_p}}^{\mathrm{S_k}}$ to be zero based on prior knowledge, e.g. set $\gamma_{\mathrm{X_1}}^{\mathrm{S_k}} = 0$ if we believe study k has the same $\mathrm{X_1}$ effect as the internal study or maybe there is not enough power to distinguish a binary $\mathrm{X_1}$ effect in study k. Similarly, if we believe the marginal effect of study k is the same as the internal study, we could force $\gamma_0^{\mathrm{S_k}}$ to be zero.

Model 4.1 is the general form of the target model as it is a saturated model allowing all intercept and possible $\mathbf{X}$ covariates to differ across populations. A special case of model 4.1 is a logistic regression model that only allows intercept differences among populations, which represents dif-

ferent prevalence and the same covariate effect in each population:

$$logit[\Pr(\mathrm{Y}=1|\mathbf{X},\mathbf{B},\mathbf{S})] = \gamma_0^{\mathrm{S}_0} + \sum_{\mathrm{k}=1}^{\mathrm{K}} \gamma_0^{\mathrm{S}_\mathrm{k}} \mathrm{S}_\mathrm{k} + \sum_{\mathrm{p}=1}^{\mathrm{P}} \gamma_{\mathrm{X}_\mathrm{p}}^{\mathrm{S}_0} \mathrm{X}_\mathrm{p} + \sum_{\mathrm{q}=1}^{\mathrm{Q}} \gamma_{\mathrm{B}_\mathrm{q}}^{\mathrm{S}_0} \mathrm{B}_\mathrm{q}. \tag{4.2}$$

We assume that the external models $\mathrm{Y}|\mathbf{X}_\mathrm{k}$'s are the best-fitted models in the class of the reduced models that was considered, but this class of reduced models may not contain the true distribution of $\mathrm{Y}|\mathrm{X}_\mathrm{k}$. One such example is when the full model is the logistic model shown in equation 4.2, but the true distribution for $\mathrm{Y}|\mathbf{X}_\mathrm{k}$'s are not logistic models as collapsibility does not hold for the logit link. We consider the fitted logistic model as the best-fitted model in the class.

4.2.2 Proposed Data Integration and Analysis Strategy

Figure 4.1 illustrates the proposed five-step strategy, along with the required assumptions in each step. We will first briefly introduce the steps and then expand upon the details.

- **Step 1:** Convert each external summary-level information into a set of synthetic data according to Section 2.2.2 in Chapter 2 and append each of the synthetic data sets to the internal data, from which we create a longer dataset as illustrated in Figure 4.1. The synthetic data for external study k constitutes of observed $\mathbf{X}_\mathrm{k}$ and the simulated value of Y. Unmeasured variables in the external populations (all $\mathbf{B}$ and some $\mathbf{X}$'s) will be treated as missing data. For example, since the external study S=1 used X_1 and X_2 to predict Y, we first replicate the observed (X_1, X_2) in S=0 a large number of times (see details of the replication number in the following descriptive paragraph for step 1); we then utilize the summary information $\mathrm{Y}|\mathrm{X}_1,\mathrm{X}_2$ from external model 1 to generate the synthetic $\hat{\mathrm{Y}}^{\mathrm{S}=1}$ values given X_1 and X_2; and lastly, the unmeasured variables X_3 and B will remain missing (Figure 4.1). Similarly for the external study S=2, we replicate the observed (X_1, X_3), and create synthetic $\hat{\mathrm{Y}}^{\mathrm{S}=2}$ values. The combined dataset is of size N $\times$ (P+Q+2).

- **Steps 2-3:** For the combined dataset created in step 1, multiply impute the missing covariates ignoring the outcome Y through multiple imputation by chained equation (MICE) to create M complete datasets, and then stack these M datasets to create a stacked dataset. These two steps are identical to step 1-2 of the stacked imputation approach proposed by Beesley and Taylor [2020], where the authors proposed to decompose the imputation model into two parts: $\mathrm{f}(\mathrm{X}_{\mathrm{miss}}|\mathrm{X}_{\mathrm{obs}},\mathrm{Y}) \propto \mathrm{f}(\mathrm{Y}|\mathrm{X})\mathrm{f}(\mathrm{X}_{\mathrm{miss}}|\mathrm{X}_{\mathrm{obs}})$ while we modify it as $\mathrm{f}(\mathrm{X}_{\mathrm{miss}}|\mathrm{X}_{\mathrm{obs}},\mathrm{B},\mathrm{Y},\mathrm{S}) \propto \mathrm{f}(\mathrm{Y}|\mathrm{X},\mathrm{B},\mathrm{S})\mathrm{f}(\mathrm{X}_{\mathrm{miss}}|\mathrm{X}_{\mathrm{obs}},\mathrm{B},\mathrm{S})$. The stacked dataset is of size MN $\times$ (P+Q+2).

- **Step 4:** Calculate weights for each row of the stacked dataset with the weights proportional to the target model density $\mathrm{f}(\mathrm{Y}|\mathbf{X}, \mathbf{B}, \mathbf{S})$. Note that all weights need to be re-scaled to 1 within individuals in the stacked dataset. Initial parameter estimates are needed for each population as discussed in subsequent paragraphs.

- **Step 5:** Estimate the parameter $\boldsymbol{\gamma}$ of the target model 4.1 through a weighted GLM using the stacked dataset. The estimated variance of $\boldsymbol{\gamma}$ can be obtained numerically through bootstrap or analytically through several existing estimators, such as the Louis information estimator or the Jackknife estimator [Beesley and Taylor, 2020, 2021].

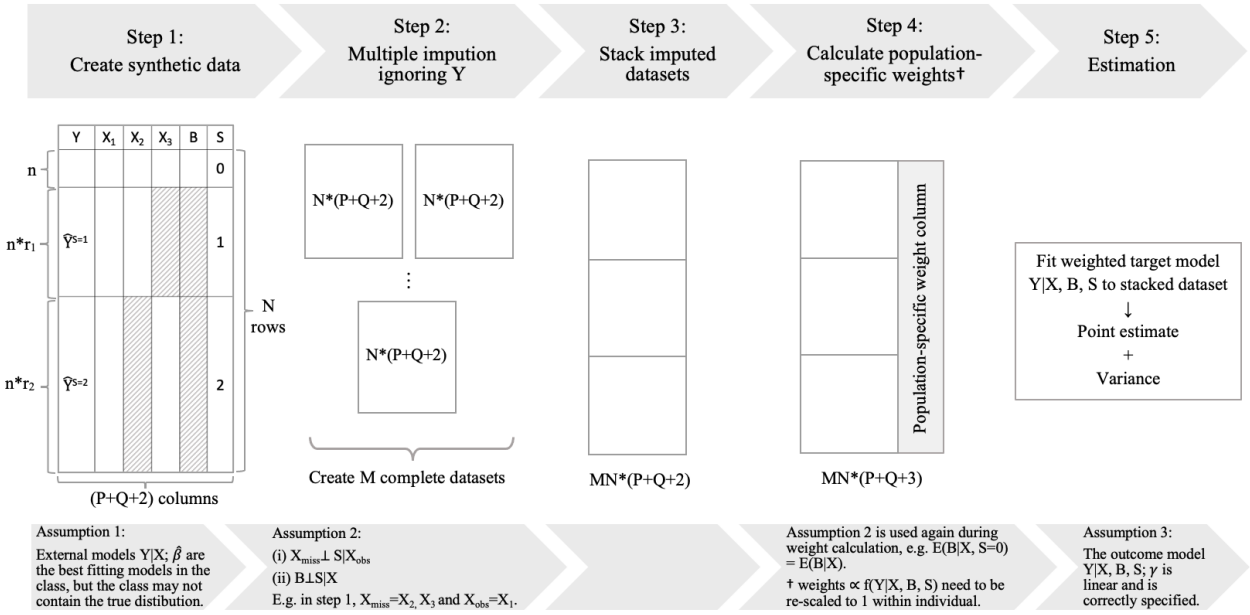


Figure 4.1: Diagram of the proposed data integration and analysis strategy.

Step 1 converts the original problem to regression modeling with missing data using a combined dataset of the internal and the synthetic data. In Chapter 2, we provided theoretical justification in special cases to show that the synthetic data method is equivalent to a constrained semi-parametric maximum likelihood approach proposed by Chatterjee et al. [2016], and it assumed the external models were the best-fitted models in the class, but the class may not contain the true distribution (Assumption 1). In finite sample size, the larger the number of replicates in each synthetic dataset (denoted as $\mathrm{r_k}$ in Figure 4.1), the more precision gain in the estimated coefficient of X; when $\mathrm{r_k}$ goes to infinity, the precision gain by incorporating external information will converge to a constant (see details in Chapter 2). In practice, it is reasonable to set the synthetic data size, i.e. $\mathrm{n}*\mathrm{r_k}$, similar to the external study's actual study size. We will assess the performance of the proposed strategy by varying the number of replicates in the simulations described in Section 4.3.

To implement steps 2 and 3, we require some quantities to be shared across populations, since the missing covariates are completely unobserved in one population, also known as block-wise missing structure. Assumption 2 contains two parts: (i) $\mathbf{X}_{\text{miss}} \perp\!\!\!\perp \text{S}|\mathbf{X}_{\text{obs}}$; and (ii) $\mathbf{B} \perp\!\!\!\perp \text{S}|\mathbf{X}$. These two assumptions imply that the conditional distribution of the missing covariates conditional on the observed covariates is the same across populations, and thus observed information can be shared across populations to impute missing covariate information. Therefore, the imputation models are $\text{f}(\text{X}_{\text{miss}}|\mathbf{X}_{\text{obs}})$ and $\text{f}(\text{B}|\mathbf{X})$ for missing $\mathbf{X}$ and missing $\mathbf{B}$, respectively (e.g. $\mathbf{X}_{\text{miss}} = [\text{X}_2, \text{X}_3]$ and $\text{X}_{\text{obs}} = \text{X}_1$ in Figure 4.1). The missing at random (MAR) assumption required by MICE is naturally satisfied as we have designed missingness, i.e., missing covariates are completely unobserved due to not being collected in the study, which is by design not related to the missing observations.

In step 4, initial parameter estimates, $\hat{\boldsymbol{\gamma}}_0$ for the internal population and $\hat{\boldsymbol{\gamma}}_{\text{k}}$'s for the external populations, are needed to calculate weights that are proportional to $\text{f}(\text{Y}|\mathbf{X}, \mathbf{B}, \text{S})$. For the internal population S=0, we replace $\text{f}(\text{Y}|\mathbf{X}, \mathbf{B}, \text{S} = 0)$ with $\text{f}(\text{Y}|\mathbf{X}, \mathbf{B}, \text{S} = 0; \hat{\boldsymbol{\gamma}}_0)$, where $\hat{\boldsymbol{\gamma}}_0$ is the internal-data-only estimates fitted on model 4.1. For external populations S=k, $\hat{\boldsymbol{\gamma}}_{\text{k}}$ from model $\text{f}(\text{Y}|\mathbf{X}, \mathbf{B}, \text{S} = \text{k}; \hat{\boldsymbol{\gamma}}_{\text{k}})$ is not directly available since we only have the summary information on the reduced model $\text{Y}|\mathbf{X}_{\text{k}}; \hat{\boldsymbol{\beta}}_{\text{k}}$. As described in Section 4.2.1, the summary information will be available in the form of either parameter estimates $\hat{\boldsymbol{\beta}}_{\text{k}}$ (Category 1) or a risk calculator of unknown form that has the ability to estimate the probability of Y=1 given $\mathbf{X}_{\text{k}}$ (Category 2). In the case of Category 1, we propose to derive the initial estimates $\hat{\boldsymbol{\gamma}}_{\text{k}} = (\hat{\boldsymbol{\gamma}}_0^{\text{S}_\text{k}}, \hat{\boldsymbol{\gamma}}_{\text{X}}^{\text{S}_\text{k}\text{T}}, \hat{\boldsymbol{\gamma}}_{\text{B}}^{\text{S}_0\text{T}})^{\text{T}}$, where $\hat{\boldsymbol{\gamma}}_0^{\text{S}_\text{k}}$ and $\hat{\boldsymbol{\gamma}}_{\text{X}}^{\text{S}_\text{k}}$ are bias-corrected estimates of intercept and $\mathbf{X}$ coefficients from $\hat{\boldsymbol{\beta}}_{\text{k}}$ according to Neuhaus and Jewell [1993] while $\hat{\boldsymbol{\gamma}}_{\text{B}}^{\text{S}_0}$ are estimated coefficients of $\mathbf{B}$ using only the internal data. Assumption 2 is used again in this step, e.g. $\text{E}(\text{B}|\text{X}, \text{S} = 0) = \text{E}(\text{B}|\text{X})$, so that we use the internal data to estimate the mean profile of $\text{B}|\mathbf{X}$ in the external populations (see Appendix C.1 for details). In the case of Category 2 where $\hat{\boldsymbol{\beta}}_{\text{k}}$ does not exist, we follow the same procedure as in Category 1 but use $\hat{\boldsymbol{\beta}}_{\text{k}}^{\text{synthetic}}$ instead. Specifically, we first create a large size of synthetic data ($\hat{\text{Y}}^{\text{S=k}}$, $\mathbf{X}_{\text{k}}^{\text{synthetic}}$) as described in Step 1, and then we fit a GLM $\hat{\text{Y}}^{\text{S=k}}|\mathbf{X}_{\text{k}}^{\text{synthetic}}$ with main effect using only the synthetic data and ignoring the missing data, from which we obtain $\hat{\boldsymbol{\beta}}_{\text{k}}^{\text{synthetic}}$. Note that this main-effect GLM is mis-specified but the best linear model in the class. Further assessment can be found in simulation II of Section 4.3.

Step 4 extends Beesley and Taylor [2020] to allow multiple populations with potentially different covariate effects in the target model. Beesley and Taylor [2020] reduces to a special homogeneous-population case of ours when all S equals to zero. In summary, we assign weights $\text{w}_{\text{im}} = \frac{\text{f}(\text{Y}_\text{i}|\text{X}_{\text{im}}, \text{B}_{\text{im}}, \text{S=k}; \hat{\boldsymbol{\gamma}}_{\text{k}})}{\sum_{\text{j}=1}^{\text{M}} \text{f}(\text{Y}_\text{i}|\text{X}_{\text{ij}}, \text{B}_{\text{ij}}, \text{S=k}; \hat{\boldsymbol{\gamma}}_{\text{k}})}$ to each observation i in the m^{th} imputed dataset, where $\text{m} \in \{1, ..., \text{M}\}$ and $\text{k} \in \{0, ..., \text{K}\}$. For internal observations, $\text{w}_{\text{im}} = \frac{\text{f}(\text{Y}_\text{i}|\text{X}_{\text{im}}, \text{B}_{\text{im}}, \text{S=0}; \hat{\boldsymbol{\gamma}}_0)}{\text{M} * \text{f}(\text{Y}_\text{i}|\text{X}_{\text{im}}, \text{B}_{\text{im}}, \text{S=0}; \hat{\boldsymbol{\gamma}}_0)} = \frac{1}{\text{M}}$.

As mentioned in Section 4.2.1, step 5 assumes that the target outcome model is linear and is correctly specified (Assumption 3). After obtaining the point estimates $\hat{\boldsymbol{\gamma}}$ by fitting the weighted

GLM, several variance estimators are available to measure the variation of $\hat{\gamma}$. Beesley and Taylor [2020, 2021] proposed three variance estimators including a bootstrap estimator by resampling the imputed datasets in step 2 and repeating steps 3-5. We propose a new bootstrap procedure by resampling the internal data and repeating all the steps 1-5, which is more computationally intense but empirically we find it gives more accurate estimation of the variance. We assess the performance of different variance estimators in simulations in Section 4.3.

4.3 Simulation Studies

Although our proposed approach can handle any target model that belongs to the class of GLM, we focus on a binary outcome and logistic regression to evaluate the performance of the proposed approach and comparison methods. In all scenarios, the internal data is of size 200, while we vary the synthetic data size from one times the internal data size (i.e. $\mathrm{r}_1 = \mathrm{r}_2 = 1$ and N=n*[$1 + \mathrm{r}_1 + \mathrm{r}_2$]=600) to 10 times (i.e. $\mathrm{r}_1 = \mathrm{r}_2 = 10$ and N=n*[$1 + \mathrm{r}_1 + \mathrm{r}_2$]=4,200) for each external population. We implement four methods, where method (1) is the benchmark, method (2) is the proposed approach, and methods (3) and (4) are two common approaches to analyze the combined dataset created in step 1 of Section 4.2.2:

1. Internal data only: fit the target model on the internal data S=0 only, without incorporating external information;

2. Proposed method: we implement it through MICE in R software. For example, in Figure 1 step 1, the imputation models for X_2, X_3 and B are $(\mathrm{X}_2|\mathrm{X}_1, \mathrm{X}_3, \mathrm{B})$, $(\mathrm{X}_3|\mathrm{X}_1, \mathrm{X}_2, \mathrm{B})$, and $(\mathrm{B}|\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3)$, respectively. Weights are calculated after imputation;

3. FCS: imputation through fully conditionally specification (FCS) by specifying an imputation model for each missing variable conditional on all the observed covariates and the outcome Y, and iteratively generate imputed values [Van Buuren et al., 2006]. For example, in Figure 1 step 1, the imputation models for X_2, X_3 and B are $(\mathrm{X}_2|\mathrm{X}_1, \mathrm{X}_3, \mathrm{B}, \mathrm{Y})$, $(\mathrm{X}_3|\mathrm{X}_1, \mathrm{X}_2, \mathrm{B}, \mathrm{Y})$, and $(\mathrm{B}|\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3, \mathrm{Y})$, respectively;

4. IMB: "imputation by ordered monotone blocks (IMB)" strategy to handle block-wise missingness proposed by Li et al. [2014]. In our case, it sequentially imputes missing covariates starting with the variable with minimum missingness conditional on the observed data, outcome, and newly imputed data. We implement IMB through MICE by specifying a different imputation model compared with FCS, e.g. in Figure 1 step 1, the imputation models for X_2, X_3 and B are $(\mathrm{X}_2|\mathrm{X}_1, \mathrm{Y})$, $(\mathrm{X}_3|\mathrm{X}_1, \mathrm{X}_2, \mathrm{Y})$, and $(\mathrm{B}|\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3, \mathrm{Y})$, respectively.

M=100 imputations are used for all multiple imputation. For FCS and IMB, we fit the same target model as the proposed method but without weights, and calculate the variance via Rubin's combining rules [Little and Rubin, 2002].

4.3.1 Simulation Settings

We provide two representative examples in Simulation I and II to illustrate how to handle the two categories of external summary-level information, respectively (Figure 4.2). Additional simulation results to assess various settings and violations of assumptions can be found in Appendix C.2.

		Simulation I*: Ideal case with continuous B1 and binary B2	**Simulation II:** External model 2 was derived by fitting a random forest model to a large dataset where the underlying generative model is a logistic regression model that contained quadratic and interaction terms
Structure of the assumed model	Internal	logit Pr[Y=1| X_1, X_2, B_1 (cont.), B_2 (binary)]	logit Pr[Y=1| X_1, X_2, X_3, X_4, B_1 (cont.), B_2 (binary)]
	External 1	logit Pr(Y=1| X_1; β_1)	logit Pr(Y=1| X_1, X_2, X_3; β_1)
	External 2	logit Pr(Y=1| X_1, X_2; β_2)	A random forest model using X_1, X_2, X_3, and X_4 to predict the probability of Y=1
Covariate distribution		(X_1, X_2, B_1) ~ N(0, 1), correlation 0.3 B_2 ~Ber[1+exp^{-1}(0.1X_1 +0.2X_2+0.3B_1)]	(X_1, X_2, X_3) ~ N(0, 1), correlation 0.3 X_4 ~N(0.1X_1 +0.1X_2+0.1X_3, 1) B_1 ~N(0.2X_1 +0.2X_2+0.2X_3+0.1X_4, 1) B_2 ~Ber[1+exp^{-1}(0.2X_1 +0.2X_2+0.2X_3+0.1X_4+0.1B_1)]
Generative true outcome model logit Pr(Y=1|X, B)	Internal	-1-X_1-X_2-B_1-B_2	-1-X_1-X_2-X_3-X_4-B_1-B_2
	External 1	1-X_1-X_2-B_1-B_2	2-X_1-X_2-X_3-X_4-B_1-B_2
	External 2	3-X_1-X_2-B_1-B_2	3-X_1-X_2-X_3-X_4-B_1-B_2+0.1(X_1^2+ X_2X_3)
Available information	Internal	Individual data (Y, X_1, X_2, B_1, B_2)	Individual data (Y, X_1, X_2, X_3, X_4, B_1, B_2)
	External 1	$\hat{\beta}_1$= (0.32, -1.19)^T †	$\hat{\beta}_1$= (1.11, -1.07, -1.16, -1.06)^T †
	External 2	$\hat{\beta}_2$= (2.11, -1.13, -1.12)^T †	A random forest risk calculator that will provide the estimated probability Y=1, given X_1, X_2, X_3, and X_4 ¶
Target model logit Pr(Y=1| X, B, S)		$\gamma_0^{S_0}$+$\gamma_0^{S_1}$$S_1$+$\gamma_0^{S_2}$ S_2 +$\gamma_{X_1}^{S_0}$$X_1$ +$\gamma_{X_2}^{S_0}$$X_2$ +$\gamma_{B_1}^{S_0}$$B_1$+$\gamma_{B_2}^{S_0}$$B_2$	$\gamma_0^{S_0}$+$\gamma_0^{S_1}$$S_1$+$\gamma_0^{S_2}$ S_2 +$\gamma_{X_1}^{S_0}$$X_1$ +$\gamma_{X_2}^{S_0}$$X_2$ +$\gamma_{X_3}^{S_0}$$X_3$ +$\gamma_{X_4}^{S_0}$$X_4$+$\gamma_{B_1}^{S_0}$$B_1$+$\gamma_{B_2}^{S_0}$$B_2$
Evaluation metrics		- Absolute bias - Variance estimation - Empirical variance of point estimates	- Area under the curve (AUC) - Sum of squared error (SSE) - Scaled Brier Score (BS)

* Additional simulation scenarios to assess the performance of the proposed approach are presented in Section 1 of Web Supplemental.
† $\hat{\beta}_k$, k=1 or 2, is estimated from a large dataset that follows the true data generating mechanism.
¶ The random forest model is fitted on a large dataset of (Y, X_1, X_2, X_3, X_4) that follows the true data generating mechanism.

Figure 4.2: Simulation settings snapshot.

- **Simulation I:** Idealized case where the internal data contains $(\mathrm{Y}, \mathrm{X}_1, \mathrm{X}_2, \mathrm{B}_1$[continuous], B_2[binary]), and two external models have been fitted to very large datasets that is sampled from the true data generating mechanism. The external models provided parameter estimates $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ from logistic regression models $\mathrm{Y}|\mathrm{X}_1$ and $\mathrm{Y}|\mathrm{X}_1, \mathrm{X}_2$, respectively. In all three populations, $\mathrm{X}_1, \mathrm{X}_2$ and B_1 follows a standard multivariate normal distribution with zero-

mean, standard deviation 1, and 0.3 correlation, while B_2 follows a Bernoulli distribution $\mathrm{B}_2|\mathrm{X}_1, \mathrm{X}_2, \mathrm{B}_1 \sim \mathrm{Ber}([1 + \exp^{-1}(0.1\mathrm{X}_1 + 0.2\mathrm{X}_2 + 0.3\mathrm{B}_1)])$. As shown in Figure 4.2, the three populations have similar generative outcome models with different intercepts, i.e. -1, 1 and 3, which give prevalence of Y=1 of 0.3, 0.57, and 0.81, respectively. The target model is a logistic regression with different intercepts of the form $\mathrm{logit}[\Pr(\mathrm{Y} = 1|\mathbf{X}, \mathbf{B}, \mathbf{S})] = \gamma_0^{\mathrm{S}_0} + \gamma_0^{\mathrm{S}_1}\mathrm{S}_1 + \gamma_0^{\mathrm{S}_2}\mathrm{S}_2 + \gamma_{\mathrm{X}_1}^{\mathrm{S}_0}\mathrm{X}_1 + \gamma_{\mathrm{X}_2}^{\mathrm{S}_0}\mathrm{X}_2 + \gamma_{\mathrm{B}_1}^{\mathrm{S}_0}\mathrm{B}_1 + \gamma_{\mathrm{B}_2}^{\mathrm{S}_0}\mathrm{B}_2$.

Evaluation metrics: We assess this simulation in terms of absolute bias, the estimated variance from bootstrap and other comparisons, and the empirical variance of point estimates.

- **Simulation II:** External model 2 was derived by fitting a random forest model to a large dataset where the underlying true generative model is a logistic regression model that contained quadratic and interaction terms. Specifically, the internal study contains complete data of $(\mathrm{Y}, \mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3, \mathrm{X}_4,\ \mathrm{B}_1$[continuous], B_2[binary]), and the two external models are available in different forms of summary information, external model 1 that provides $\hat{\boldsymbol{\beta}}_1$ from a logistic regression model $\mathrm{Y}|\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3$ and external model 2 that can provide the estimated probabilities of Y=1 given $\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3$ and X_4 through a fitted random forest model. In all three populations, $\mathrm{X}_1, \mathrm{X}_2$ and X_3 follows a standard multivariate normal distribution with zero-mean, standard deviation 1, and 0.3 correlation, while X_4 and B_1 each follows a conditional normal distribution, $\mathrm{X}_4|\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3 \sim \mathrm{N}(0.2\sum_{\mathrm{p}=1}^{3}\mathrm{X}_\mathrm{p}, 1)$ and $\mathrm{B}_1|\mathbf{X} \sim \mathrm{N}(0.2\sum_{\mathrm{p}=1}^{3}\mathrm{X}_\mathrm{p} + 0.1\mathrm{X}_4, 1)$, and B_2 follows a Bernoulli distribution $\mathrm{B}_2|\mathbf{X}, \mathrm{B}_1 \sim \mathrm{Ber}\big(\{1 + \exp^{-1}[0.2\sum_{\mathrm{p}=1}^{3}\mathrm{X}_\mathrm{p} + 0.1(\mathrm{X}_4 + \mathrm{B}_1)]\}\big)$, respectively. Similar to simulation I, the true generative distributions of Y in the internal and external population 1 shared the same main covariate effect but have different intercepts (-1 and 2, which corresponds to prevalence 0.3 and 0.65), while external model 2 additionally contains a quadratic term and an interaction (with intercept 3 that corresponds to prevalence 0.73). The target model is a logistic regression with the form $\mathrm{logit}[\Pr(\mathrm{Y} = 1|\mathbf{X}, \mathbf{B}, \mathbf{S})] = \gamma_0^{\mathrm{S}_0} + \gamma_0^{\mathrm{S}_1}\mathrm{S}_1 + \gamma_0^{\mathrm{S}_2}\mathrm{S}_2 + \gamma_{\mathrm{X}_1}^{\mathrm{S}_0}\mathrm{X}_1 + \gamma_{\mathrm{X}_2}^{\mathrm{S}_0}\mathrm{X}_2 + \gamma_{\mathrm{X}_3}^{\mathrm{S}_0}\mathrm{X}_3 + \gamma_{\mathrm{X}_4}^{\mathrm{S}_0}\mathrm{X}_4 + \gamma_{\mathrm{B}_1}^{\mathrm{S}_0}\mathrm{B}_1 + \gamma_{\mathrm{B}_2}^{\mathrm{S}_0}\mathrm{B}_2$.

  As described in step 4 of Section 4.2.2, $\hat{\boldsymbol{\beta}}_1$ can be directly used to calculate weights for S=1 while we need to estimate $\boldsymbol{\beta}_2^{\mathrm{synthetic}}$ to calculate weights for S=2. To obtain $\boldsymbol{\beta}_2^{\mathrm{synthetic}}$, we first generate a large synthetic data set $(\hat{\mathrm{Y}}^{\mathrm{S}=2}, \mathrm{X}_1^{\mathrm{synthetic}}, ..., \mathrm{X}_4^{\mathrm{synthetic}})$ by replicating the observed $(\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3, \mathrm{X}_4)$ and generating $\hat{\mathrm{Y}}^{\mathrm{S}=2}$ values through the available random forest model, and then fit a main effect logistic model $\hat{\mathrm{Y}}^{\mathrm{S}=2}|\mathrm{X}_1^{\mathrm{synthetic}}, ..., \mathrm{X}_4^{\mathrm{synthetic}}$ using only the synthetic data and ignoring the missing B_1 and B_2.

  Evaluation metrics: Since prediction accuracy will be the main goal in such situation in practice, we evaluate this simulation using three prediction metrics over a validation data of size $\mathrm{N}_{\mathrm{test}} = 2,000$: Area under the curve (AUC); Sum of squared error (SSE) $= \frac{1}{\mathrm{N}_{\mathrm{test}}}\sum_{\mathrm{i}=1}^{\mathrm{N}_{\mathrm{test}}}(\hat{\mathrm{p}}_\mathrm{i} -$

$p_{i0})^2$, where $\hat{p}_i$ and p_{i0} denotes the estimated and true probability of $Y_i = 1$ given $\mathbf{X}_i$ and $\mathbf{B}_i$, respectively; and Scaled Brier Score (BS): $= \sum_{i=1}^{N_{test}}(Y_i - \hat{p}_i)^2 / \sum_{i=1}^{N_{test}}(Y_i - \bar{Y})^2$, where $\bar{Y} = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} Y_i$.

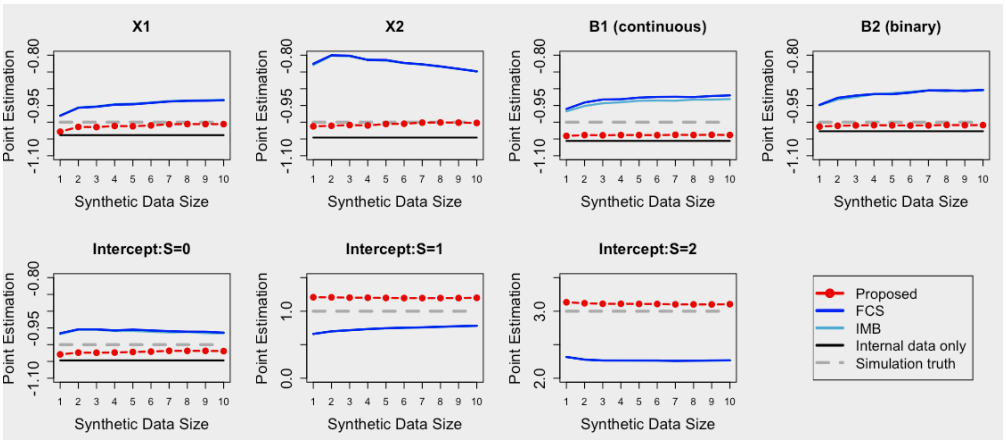
4.3.2 Simulation Results

Figure 4.3 shows the average results of the target model parameter estimates across 500 simulated datasets for simulation I, including point estimates in Figure 4.3a, variance estimators versus the empirical variance of the point estimates in Figure 4.3b, and the comparison of different variance estimators for the proposed strategy in Figure 4.3c. This figure appears in color in the electronic version of this article, and color refers to that version. Figure 4.3a shows that FCS (dark blue curve) and IMB (light blue curve) have similarly biased estimates, indicating these traditional imputation strategies can not distinguish heterogeneous population effects, while the proposed method (red dotted curve) always shows close results to the truth (grey dashed curve) for all covariates, especially X_2 and external intercepts S=2 where other methods show severe bias. For example, the absolute bias of X_2 coefficient estimates can be up to 0.2 for both FCS and IMB while it's only 0.01 for the proposed method.

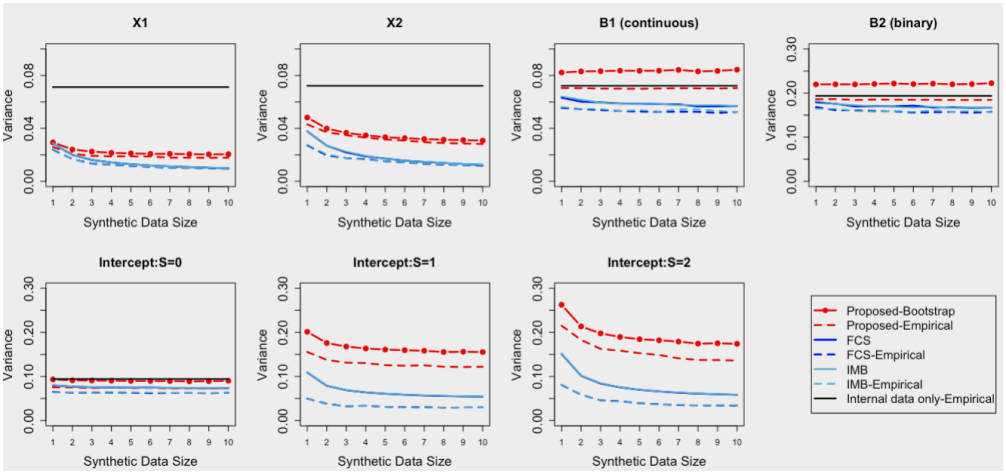
As shown in Figure 4.3b, each color denotes one distinct method, along with one solid curve represents the variance estimator, and one dashed curve represents the Monte Carlo empirical variance of the point estimates. If the variance is correctly estimated, the solid curve should be approximately equal to the corresponding dashed one, which is true for all methods. As expected, all methods show precision gain in estimated X coefficients compared to the internal data only (the longer the distance to the black internal-data-only curve, the larger the precision gain) while no precision gain is found in B covariates and the intercept due to no external added information and allowing population-specific effects, respectively. The proposed method has over 50% efficiency gain in estimated X coefficients compared to the internal data. We see FCS and IMB have larger precision gain in both estimated X coefficients than the red proposed curve, which may be explained by bias-variance trade-off as they also have larger bias in the corresponding point estimates in Figure 4.3a. We will discuss the underlying statistical reason in Section 4.5.

As shown in Figure 4.3c, we show the result of the Louis information estimator (StackImpute-Louis), one of the three variance estimators proposed by Beesley and Taylor [2020] as they always have similar performances. We have shown in Chapter 2 that when the synthetic data size goes to infinity, the precision gain we achieve in X covariates will converge to a constant, which is shown as the gradually stable trend of the grey curve (Monte Carlo empirical variance of the point estimates, also serves as the empirical truth here). When the synthetic data size increases from one times the internal data size to 10 times for each external study (i.e. total missing rate increases from 66.6%

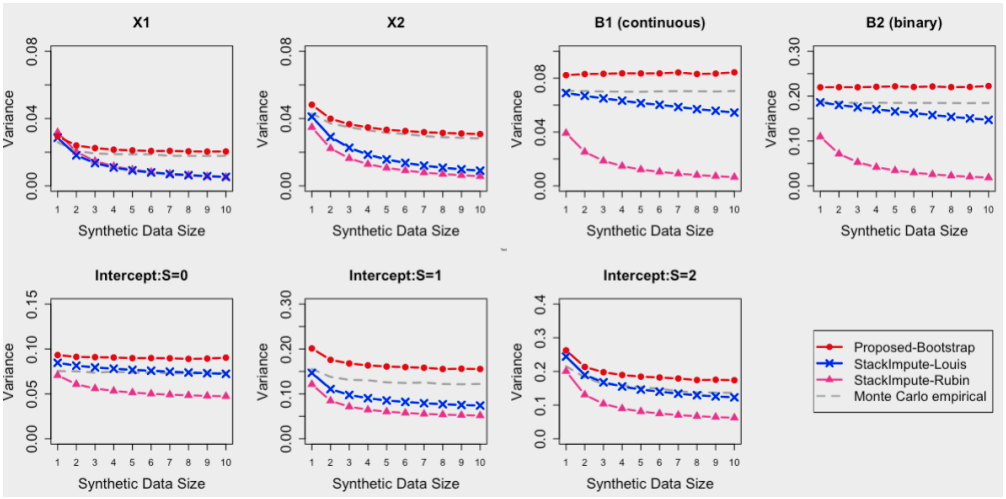
to 95%), the StackImpute variance estimator and Rubin's rule variance continuously underestimate the empirical truth. On the contrary, the proposed variance estimator by bootstrapping the whole proposed procedure is always close to the empirical truth, especially in X covariates where the bias in other methods can be 10 times higher than the proposed method (i.e. 0.02 versus 0.002 in absolute bias) and could be even larger when the synthetic data size keep increasing. Moreover, in estimating the variation of the coefficient corresponding to B_2, the proposed method has stable bias around 0.035 while the other methods exhibit substantially larger bias. Note that the internal-data-only results (black solid curve) does not exist in external intercepts as they were not available in the internal data, whereas the bias of the internal data estimates is due to the small sample bias compared with the simulation truth.



(a) Point estimates



(b) Variance estimators vs. the empirical variance



(c) Different variance estimators of the proposed method

Figure 4.3: Visualization of simulation I results over increasing synthetic data size (a) point estimates (b) variance estimators vs. the empirical variance (c) different variance estimators of the proposed strategy.

Figure 4.4 shows the performance of each method in Simulation II over increasing synthetic data size on a validation dataset that follows the true data generating mechanism, with three prediction metrics on the row and three populations on the column. In general, the results in Figure 4.4 are in line with Figure 4.3a, implying that the proposed method has better overall prediction performance compared with others. Specifically, in the first column (internal population S=0), all methods incorporating external information have consistently better prediction ability (larger AUC, smaller SSE and smaller BS) than using the internal data only (the dashed grey line). While all methods have similar performance in terms of AUC (first row) and predicting internal population (first column), the proposed approach outperforms others in terms of SSE and BS in predicting external populations, especially external study 2 where the true parameter values are quite different from the internal study values (the proposed method has up to 41% more improvement in SSE and 19% more improvement in BS compared with FCS and IMB). The proposed method shows a modest improvement in performance as the size of the synthetic data increases, e.g. for S=2, SSE decreases 12.9% from 9.3 to 8.1 in the proposed method when the synthetic data size increases from one to 10. In this particular scenario, there is little gain in performance with synthetic data more than 4 times the internal data size. Note that it is hard to distinguish FCS and IMB in the figure as they have very close results.

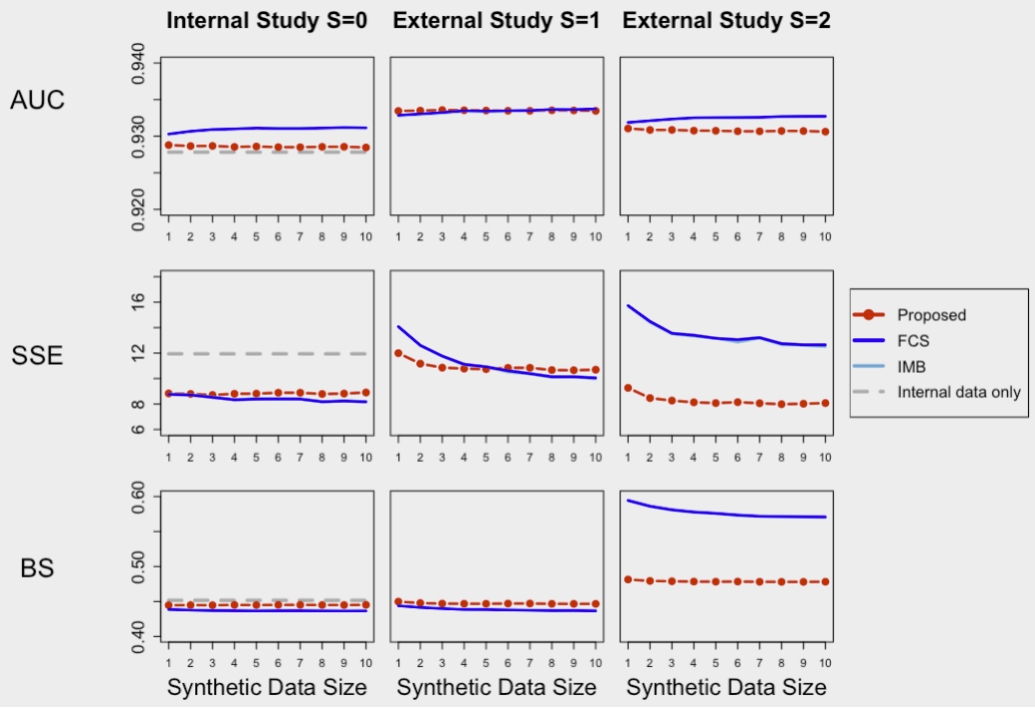


Figure 4.4: Visualization of prediction metrics over increasing synthetic data size for simulation II. Larger AUC, smaller SSE and smaller BS represents better prediction.

4.4 Application to Prostate Cancer Data

We apply the proposed method to predict the risk of high-grade prostate cancer (Gleason score over 6) using an internal dataset containing patients from three United States academic institutions [Tomlins et al., 2015] and two external risk calculators, one Prostate Cancer Prevention Trial risk calculator established from a United States population [Thompson et al., 2006] and another European Randomized Study of Screening for Prostate Cancer risk calculator 3 established from a European population [Roobol et al., 2012].

The external risk calculators each used slightly different predictors to predict the same outcome through logistic regression, which we will denote as PCPThg and ERSPC, respectively. PCPThg and ERSPC both used prostate-specific antigen level (PSA) and digital rectal examination findings (DRE) as one of the predictors, and PCPThg also used age, race (African American or not) and prior biopsy results while ERSPC additionally used transrectal ultrasound prostate volume (TRUS-PV):

- PCPThg: $\text{logit}(\text{p}_\text{i}) = -3.69 + 0.89\log_2(\text{PSA}_\text{i}) + \text{DRE}_\text{i} + 0.03\text{Age}_\text{i} + 0.96\text{Race}_\text{i} - 0.36\text{Biopsy}_\text{i}$;
- ERSPC: $\text{logit}(\text{p}_\text{i}) = -3.16 + 1.18\log_2(\text{PSA}_\text{i}) + 1.81\text{DRE}_\text{i} - 1.51\log_2(\text{TRUS-PV}_\text{i})$,

where p_i is the probability of observing high-grade prostate cancer for subject i. A more detailed description of the above two equations can be found in Appendix B.5 of Chapter 3.

In the internal individual-level data, in addition to all the predictors used in PCPThg and ERSPC, we also have data on two new biomarkers, prostate cancer antigen 3 (PCA3) and TMPRSS2:ERG (T2:ERG) gene fusions, that were prognostic of prostate cancer [Tomlins et al., 2015, Truong et al., 2013]. Therefore, in Table 4.1, we present the results of two target models using a total of eight predictors, including these two new biomarkers, one model only allows the intercept to be different across populations ("different intercept only" model corresponds to model 4.2), and another flexible model allows all possible covariates that used in the external populations to have population-specific effects ("different intercept and covariates" model that corresponds to model 4.1). The detailed model forms can be found in the table legend. A total of 678 male patients who had complete data were included in the internal data set provided by Tomlins et al. [2015], and an additional 1,174 patients' data were independently collected from seven community clinics throughout the United States for validation.

The grey empty blocks in Table 4.1 imply that these predictors were not used in the certain external model, and thus in the proposed method we assume they have the same coefficient as the internal population (grey blocks with values). Results in Table 4.1 show (i) we will gain precision of the estimated coefficients by incorporating external model information (smaller SE highlighted in green), e.g. in the different-intercept-only model, the bootstrap SE of $\log_2$(PSA) reduces from

Table 4.1: Results of the data example for predicting the risk of high-grade prostate cancer. Internal dataset of size n=678; validation dataset of size N_{test}=1,174; SE, standard error; green represents good performance with large precision gain or better model calibration, grey represents certain predictor was not used in the external calculator, yellow represents population-specific effect different from the internal population, and red represents poor performance compared with direct regression; SE in the proposed method are (StackImpute SE) [bootstrap SE from 500 replicates].

		PCPThg		ERSPC			Proposed method					
		Original	Estimated	Original	Estimated	Direct regression	Different intercept only†			Different intercept and covariates ‡		
							Internal	PCPThg	ERSPC	Internal	PCPThg	ERSPC
Point estimates (SE)	Intercept	-3.686 /	-1.409 (.115)	-3.16 /	-1.391 (.111)	-4.202 (.455) [REF]	-4.157 (.421) [.456]	-6.884 (.464) [.516]	-6.013 (.449) [.515]	-4.178 (.432) [.483]	-6.884 (.478) [.531]	-5.953 (.464) [.504]
	$\log_2$(PSA)	0.894 /	0.735 (.124)	1.176 /	0.891 (.124)	0.885 (.146) [REF]	1.082 (.060) [.070]	1.082 (.060) [.070]	1.082 (.060) [.070]	0.877 (.144) [.152]	1.099 (.091) [.101]	1.202 (.097) [.100]
	DRE	1 /	1.145 (.257)	1.813 /	1.306 (.269)	1.045 (.299) [REF]	1.118 (.117) [.139]	1.118 (.117) [.139]	1.118 (.117) [.139]	1.044 (.298) [.310]	0.707 (.182) [.229]	1.614 (.179) [.184]
	Age	0.03 /	0.033 (.012)			0.032 (.014) [REF]	0.030 (.008) [.010]	0.030 (.008) [.010]	0.030 (.008) [.010]	0.032 (.013) [.014]	0.030 (.010) [.012]	0.032 (.013) [.014]
	Biopsy	-0.36 /	-1.444 (.272)			-1.201 (.294) [REF]	-0.618 (.150) [.184]	-0.618 (.150) [.184]	-0.618 (.150) [.184]	-1.195 (.291) [.303]	-0.069 (.181) [.207]	-1.195 (.291) [.303]
	Race	0.96 /	0.442 (.288)			0.183 (.331) [REF]	0.434 (.175) [.222]	0.434 (.175) [.222]	0.434 (.175) [.222]	0.186 (.326) [.360]	0.680 (.205) [.258]	0.186 (.326) [.360]
	$\log_2$ (TRUS-PV)			-1.514 /	-1.697 (.225)	-1.699 (.254) [REF]	-1.553 (.145) [.175]	-1.553 (.145) [.175]	-1.553 (.145) [.175]	-1.656 (.243) [.271]	-1.656 (.243) [.271]	-1.443 (.579) [.611]
	$\log_2$ (PCA3+1)					0.495 (.089) [REF]	0.486 (.083) [.092]	0.486 (.083) [.092]	0.486 (.083) [.092]	0.494 (.084) [.094]	0.494 (.084) [.094]	0.494 (.084) [.094]
	$\log_2$ (T2:ERG+1)					0.098 (.038) [REF]	0.092 (.036) [.037]	0.092 (.036) [.037]	0.092 (.036) [.037]	0.094 (.037) [.039]	0.094 (.037) [.039]	0.094 (.037) [.039]
Prediction metrics**	AUC	0.700	0.707	0.720	0.723	0.806 (REF)	0.804	0.804	0.804	0.805	0.797	0.802
	Scaled Brier score	1.059	0.987	0.954	0.994	0.903 (REF)	0.915	0.977	0.875	0.902	0.990	0.875

† $\gamma_0^{S_0} + \gamma_0^{S_{PCPThg}} S_{PCPThg} + \gamma_0^{S_{ERSPC}} S_{ERSPC} + \gamma_{X_1}^{S_0} \log_2(PSA_i) + \gamma_{X_2}^{S_0} DRE_i + \gamma_{X_3}^{S_0} Age_i + \gamma_{X_4}^{S_0} Biopsy_i + \gamma_{X_5}^{S_0} Race_i + \gamma_{X_6}^{S_0} \log_2(\text{TRUS-PV}_i) + \gamma_{B_1}^{S_0} \log_2(PCA3_i + 1) + \gamma_{B_2}^{S_0} \log_2(\text{T2:ERG}_i + 1)$ (*)

‡ (*) + $S_{PCPThg}[\gamma_1^{S_{PCPThg}} \log_2(PSA_i) + \gamma_2^{S_{PCPThg}} DRE_i + \gamma_3^{S_{PCPThg}} Age_i + \gamma_4^{S_{PCPThg}} Biopsy_i + \gamma_5^{S_{PCPThg}} Race_i] + S_{ERSPC}[\gamma_1^{S_{ERSPC}} \log_2(PSA_i) + \gamma_2^{S_{ERSPC}} DRE_i + \gamma_6^{S_{ERSPC}} \log_2(\text{TRUS-PV}_i)]$

**prediction metrics on validation dataset

0.146 to 0.070 while it reduces for DRE from 0.299 to 0.139, compared with direct regression; (ii) when allowing population-specific effects (yellow blocks), we will not expect to see much precision gain due to variance-bias trade-off, e.g. both precision gain of $\log_2$(PSA) and DRE we see in the different-intercept-only model diminishes in the different-intercept-and-covariates model; and (iii) similar to the results in simulations, the analytical SE (in the round parenthesis) tends to provide a smaller estimation than the bootstrap SE (in the squared parenthesis) and potentially underestimates the true variability of the estimates.

In the prediction metrics row, we show AUC (higher value represents better discrimination) and scaled Brier score (smaller value means better calibration) calculated using the validation cohort, where the red blocks implies slightly worse overall predictive performance of PCPThg population compared to direct regression, e.g. 0.9% reduction of AUC. This may be because that the validation cohort may represent a moderately different population than the training cohort as it has different baseline distribution as noted by Tomlins et al. [2015], which may also explain why the fitted model for the European population has better performance on the validation data.

4.5 Discussion

Flexibility in external models and populations: The proposed approach adds to the existing research on integrating external summary information into the internal study. It can develop improved models and provide statistical inference not only for the internal population but also for the external populations. The parameters of the external population models are allowed to differ from those of the internal population. This new strategy has the appealing feature of being able to make use of external information that comes in the form of a "black box" algorithm, i.e. an algorithm that provides a predicted probability, but the underlying model is not necessarily simple or transparent or even known. The key aspect is that the external information allows the creation of synthetic data. We further summarize some key points and concluding remarks of the proposed strategy in the subsequent paragraphs.

Using partial information through data integration: The proposed method can integrate summary information from multiple external models that each uses different covariates $\mathbf{X}_{\mathrm{k}} \in \mathbf{X}$ into the current study. The simulation and real data analysis showed expected results that we can only gain precision on the estimated coefficient of $\mathbf{X}$ but not the $\mathbf{B}$ coefficients that are only available internally, even when $\mathbf{X}$ and $\mathbf{B}$ are correlated. This is consistent with the theoretical results in Dai et al. [2012] that the MLE estimator $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}_{\mathrm{B}}$ are always asymptotically independent under regularity conditions, where $\hat{\boldsymbol{\beta}}$ is the estimates of intercept and $\mathbf{X}$ coefficients in model $\mathrm{Y}|\mathbf{X};\boldsymbol{\beta}$ and $\hat{\boldsymbol{\gamma}}_{\mathrm{B}}$ is the estimated coefficient of $\mathbf{B}$ from model $\mathrm{Y}|\mathbf{X},\mathbf{B};\boldsymbol{\gamma}$, respectively.

Principled inference post-imputation: Since the proposed strategy uses the stacked multiple

imputation (StackImpute) proposed by Beesley and Taylor [2020], it also borrows strengths from StackImpute to avoid incompatibility between the imputation model and the analysis model, and can accommodate complicated outcomes such as the time-to-event outcome in survival models. In the proposed method, we introduced two types of variance estimation, the proposed bootstrap variance and the analytical variance estimators proposed in Beesley and Taylor [2020, 2021]. In general, the bootstrap variance can provide more valid variance estimation but may be more computationally intense compared with others, while analytical estimates are fast to compute but may be biased. Based on simulation results, when the predictors have small covariate effects (simulation in Appendix C.2.2), the bias of the analytical variance estimates is small.

Improve not just current study but external model predictions: To our knowledge, very few existing approaches can allow different population effects through regression analysis in data fusion, let alone improving external model predictions. It is worth noting that this same problem setting has a wide range of applications beyond regression analysis. Several approaches are proposed in the causal inference field to estimate the average causal effect, aiming to incorporate the supplementary information from the validation dataset to the main dataset and allowing heterogeneous treatment effects among different data sources. For example, Antonelli et al. [2017] proposed a unified Bayesian imputation framework built upon the work by Wang et al. [2012b], introducing a dependence parameter to represent the prior odds of including a predictor in the outcome model given that it is in the exposure model, and assuming different population indicators in the outcome model. Yang and Ding [2020b] posited a stochastic framework on the estimators from different data sources which flexibly leverages the supplementary information from validation datasets and they used a sensitivity parameter to quantify the systematic difference among data sources. Similarly, Huang and Qin [2020] and Chen et al. [2020] addressed this same problem in the implementation of survival data without assuming comparability among data sources.

Allow for violation of transportability assumption: The transportability assumption is common in data integration and causal inference when certain variables are not mutually observed across populations [Rassler, 2004, Reiter, 2012, Bareinboim and Pearl, 2013]. Compared with more strict transportability assumed in the literature, e.g. Chatterjee et al. [2016] assumed transportability of the joint distribution of Y, X, B while Antonelli et al. [2017] assumed conditional transportability of $\mathrm{X}_{\mathrm{miss}}|\mathrm{X}_{\mathrm{obs}}, \mathrm{Y}$, we only require conditional transportability among covariates (Assumption 2). While the simulation results (Appendix C.2.4) suggest that violating this assumption could have a mild impact, one can consider applying additional shrinkage methods such as the empirical Bayes approaches proposed by Estes et al. [2017] and us (in Chapter 3) after obtaining estimates from the proposed approach, which can empirically strike a balance between bias and efficiency when the transportability between populations is unclear.

Limitations: While the results of the simulation study suggest that the proposed strategy has

promising performance in providing both accurate statistical inference and prediction compared with comparison methods, some limitations are worth noticing. Particularly, the proposed method relies on good initial estimates for each external population. We propose to use a geometric approach by utilizing the observed data relationship to map the parameter estimates in the reduced model (i.e. $\mathrm{Y}|\mathrm{X};\boldsymbol{\beta}$) to the target model (i.e. $\mathrm{Y}|\mathrm{X},\mathrm{B};\boldsymbol{\gamma}$). While the simulation results show promising performance, caution must be exercised during implementation when the true underlying relationship is hard to verify. In a special case where the external study has the same population distribution as the internal population, the internal data estimates can directly serve as the initial estimates for the external population.

Future directions: An interesting extension of the proposed method is to accommodate the situation where selection bias exists and selection probability or survey weights are available for each observation in the internal population. In theory, the proposed method can be adapted to accommodate this by replacing the synthetic Y values with the inverse probability-weighted or survey weights-weighted synthetic Y values in step 1 of the proposed strategy. Alternatively, instead of copying the whole internal X's multiple times to create the same X distribution as the internal population, one can consider proportionally creating synthetic X through the given weights to recovery the representative distribution in the external populations. Further investigation is needed to evaluate this. Furthermore, if the exposure indicator is available as a covariate in all populations, one can also use the regression estimates from the proposed method to calculate the estimated average causal effect by averaging over the joint distribution of $(\mathbf{X},\mathbf{B})$. On the contrary, it is unclear whether we can directly use the intermediate parameter estimates from causal inference methods for regression inference. For example, the causal inference approach–guided Bayesian method adjusting for unmeasured confounding [Antonelli et al., 2017] aimed at obtaining unbiased causal effects by averaging selective regression models, can also produce the regression estimates for all covariates, which is the same as our goal. We attempted to modify their method and code to serve our purpose but the results did not seem promising, this may have been because a direct comparison of the performance of two approaches is not appropriate when they have different goals.

4.6 Software and Publication

R package SynDI implementing the proposed method can be found on GitHub at https://github.com/umich-biostatistics/SynDI. The content of this chapter has been submitted for publication. A preprint is available on arXiv at https://arxiv.org/abs/2106.06835.

CHAPTER 5

Discussion

In this dissertation, we propose three statistical methods in Chapter 2, 3 and 4, respectively, to incorporate external summary-level information into the regression analysis of a current study. In Chapter 2, we propose a novel synthetic data method that can convert the external model information into synthetic data, where the only required external model information is the ability to generate predictions of the outcome given covariates. We introduce the synthetic data method under the setting of one single external study. In Chapter 3, we propose a meta-inference framework using an empirical Bayes (EB) approach considering multiple external studies, where each of the external models provides a set of regression estimates. The meta-inference framework is efficient and robust to make valid inference of the internal study as it selects the compatible external model estimates for the internal study and adaptively assigns weights according to compatibility. In Chapter 4, we extend the synthetic data method to accommodate the situation with multiple external models and further allow for heterogeneous covariate effects across external populations.

Different approaches to solve the same genre of problem: The chapters of this dissertation form a coherent whole, with the consistent goal to solve the same data integration problem but each has different priorities. All three proposed methods are flexible to incorporate external models that use a slightly different set of covariates, as long as the covariates are the subset of the observed covariates in the internal study. The synthetic data method proposed in Chapter 2 relaxes the requirements of the information that is available externally compared with the traditional constrained maximum likelihood approaches. In Chapter 3, the meta-inference framework focuses on making robust inference for the internal study, i.e. selecting the most comparable information to the internal study when multiple external models are available for use, aiming to strike a balance between efficiency gain and making valid statistical inference for the internal study. In Chapter 4, in addition to the features we captured in the previous two chapters (i.e. leveraging auxiliary information from a broad class of externally fitted model or established risk calculators of unknown form as in Chapter 2; and accommodating multiple external model information in Chapter 3), we further allow heterogeneous covariate effects across the external populations by extending the synthetic data method proposed in Chapter 2.

In summary, (i) Chapter 2 considers one single external study while Chapter 3 and 4 consider the situation where summary-level information from multiple external studies are available for use; (ii) Chapter 2 and 3 focus on making improved inference for the internal study while Chapter 4 can also make inference for the external populations; (iii) Chapter 2 and 4 can handle more flexible external information, from parametric regression model estimates to any machine learning models of unknown form, compared with Chapter 4, where regression model estimates are needed; and (iv) due to EB estimator's robust feature from shrinkage effect, Chapter 3 has less requirement on similarities across populations (i.e. collapsibility and transportability assumptions) compared to others, which we will discuss in details in the subsequent paragraph.

Transportability and collapsibility assumptions in data integration: The issues of transportability of distributions and collapsibility of prediction models are critical and inevitable in data integration. Transportability concerns the distributional similarity across the populations while collapsibility relates to the models being used within the population. In this dissertation, collapsibility is concerned with whether the distribution implied by the $\mathrm{Y}|\mathbf{X}_\mathrm{k}$ model is compatible with the distribution implied by the $\mathrm{Y}|\mathbf{X},\mathbf{B}$ model. In Chapter 2 and 4, the collapsibility assumption asks for that the external models $\mathrm{Y}|\mathbf{X}_\mathrm{k}$'s were the best-fitted models in the reduced class that was considered, but this class of reduced models might not contain the true distribution of $\mathrm{Y}|\mathbf{X}_\mathrm{k}$. Full or partial transportability assumption (i.e. transportability of the joint distribution Y, X, B or conditional transportability of $\mathrm{Y}|\mathrm{X},\mathrm{B}$ distribution) is commonly assumed in the literature when certain variables are not mutually observed across populations [Antonelli et al., 2017, Chatterjee et al., 2016, Estes et al., 2017] yet hard to verify in practice. In Chapter 4, we require mild conditional transportability among covariates X and B for valid imputation, violating which had a mild impact as shown in simulations. No specific transportability or collapsibility assumption is needed in Chapter 3 as the EB estimator's shrinkage effect limits the impact of external model information that is not compatible with the internal data and thus protects against the severe bias. Another advantage of the proposed methods is that in all three methods, we do not require the marginal $\mathbf{X}$ distribution to be the same across populations, as is assumed in other literature [Chatterjee et al., 2016, Kundu et al., 2019].

In practice, one can consider applying additional shrinkage methods such as the EB approaches proposed by Estes et al. [2017] or in Chapter 3 after obtaining estimates from other proposed approaches, which can empirically strike a balance between bias and efficiency when the transportability between populations is unclear. However, as we saw in simulations, the efficiency will decrease in order to trade for some level of robustness.

Gaining efficiency by using partial information through data integration: The proposed methods can integrate summary information from single or multiple external models that each uses different covariates $\mathbf{X}_\mathrm{k} \in \mathbf{X}$ into the current study. The simulations and real data analysis

in all three chapters show that we can gain around 10% to over 50% efficiency on the estimated coefficient of $\mathbf{X}$ compared with the internal-data-only analysis, where the factors that limit the efficiency gain include that too few external models used certain covariate, the internal sample size is large so that the largest possible improvement is small, or the shrinkage method is applied to protect the potential bias. When we evaluate the prediction metrics such as SSE, the improvement compared with others is around 16%–40% while the magnitude of improvement in AUC or Brier score is relatively small around 1%–20%. We also notice an expected result that we can only gain precision on the estimated coefficient of $\mathbf{X}$ but not the $\mathbf{B}$ coefficients that are only available internally, even when $\mathbf{X}$ and $\mathbf{B}$ are correlated. This is consistent with the theoretical results in Dai et al. [2012] that the MLE estimator $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}_{\mathrm{B}}$ are always asymptotically independent under regularity conditions, where $\hat{\boldsymbol{\beta}}$ is the estimates of intercept and $\mathbf{X}$ coefficients in model $\mathrm{Y}|\mathbf{X};\boldsymbol{\beta}$ and $\hat{\boldsymbol{\gamma}}_{\mathrm{B}}$ is the estimated coefficient of $\mathbf{B}$ from model $\mathrm{Y}|\mathbf{X},\mathbf{B};\boldsymbol{\gamma}$, respectively.

Novelty of the synthetic data method: It is worthy to highlight the advantages of the synthetic data method we proposed in Chapter 2. By creating large pseudo-data that is compatible with the externally established model, the synthetic data method naturally incorporates the external summary-level information into the internal data. Compared with converting the external information into constraint as in CML approaches, the synthetic data method not only simplified the task from solving complex constrained optimization, but also provides a potentially more flexible and general framework to handle this problem. Any tools that can handle missing data, including the multiple imputation technique we used, can be applied to analyze the combined dataset of the internal and the synthetic data. The only requirement for the synthetic data approach is the ability to generate outcome values given covariates from the information of the external models, without the need to know the exact form of the model. The extension to multiple external populations in Chapter 4 further allows researchers to make statistical inference on both the internal and the external populations. It is broadly applicable for general data types for the outcome and covariates, when covariates are of multi-dimension, and when each of the external models uses a different subset of covariates. These features are particularly appealing when both the parametric regression modeling and the machine learning algorithms with non-trivial form have been widely used in risk prediction modeling.

Size of the internal and external studies: In this dissertation, we assume the external models are well-established on large data that produce credible estimates or predicted outcomes. In the synthetic data methods (Chapter 2 and 4), the theoretical results suggest creating a very large size of synthetic data for each external study in order to gain the maximum possible efficiency gain on the condition that the external model $\mathrm{Y}|\mathbf{X}_{\mathrm{k}}$ is compatible with the target model $\mathrm{Y}|\mathbf{X},\mathbf{B}$. In practice when the exact compatibility is hard to satisfy, we recommend limiting the synthetic data size similar to the external study's actual study size. In Chapter 3, we provide an option to account

for the external model uncertainty by taking into account the variance-covariance matrix of the external model estimates.

Design of the internal and external studies: We have considered various sources of information variation across the models and populations in this dissertation, e.g. different forms of external model information, different subsets of covariates used by each of the external models, and heterogeneous covariate effects across populations. Caution must be exercised as there could still be fundamental variability due to design and sampling differences across studies. Examples of these include case-control studies, outcome-dependent sampling, selection bias in who is included in the dataset (either internal or external) and measurement bias during data collection, ignoring which could lead to biased inference in data integration.

Future directions: One possible extension of the proposed method is the application in causal inference. In all three chapters, regression estimates of the target model are our ultimate product. If the treatment indicator was available as one of the X covariates, we could directly calculate the estimated average causal effect by averaging over the joint distribution of $(\mathbf{X}, \mathbf{B})$ through the formula $\mathrm{E}_{\mathrm{X,B}}[\mathrm{E}(\mathrm{Y}|\mathrm{treatment} = 1, \mathbf{X}, \mathbf{B}) - \mathrm{E}(\mathrm{Y}|\mathrm{treatment} = 0, \mathbf{X}, \mathbf{B})]$ using the regression estimates obtained from the proposed methods.

Another interesting extension is to modify the proposed methods to accommodate electronic health record (EHR) data under a slightly different setting and goal. Instead of having a moderately-sized unbiased internal dataset and several external models fitted on large data as in the current setting, we would then consider having a massively large number of internal EHR data, which could provide us potentially biased results due to selection bias, a common issue in EHR data. The goal would become to correct the bias from EHR data by incorporating the external information from models that are fitted on smaller sample sizes (compared to the large size of EHR data).

Another point of future consideration is to derive the asymptotic variance estimates of the target point estimates in Chapter 4 (i.e. the point estimates from the weighted GLM using the stacked dataset after multiple imputation). We showed in simulations that the existing fast-to-calculate variance estimators [Beesley and Taylor, 2020, 2021] only had unbiased results when the covariate effect is small or the synthetic data size is small (e.g. one times the internal data size), which contradicts the fact that larger synthetic data size can guarantee the efficiency gain. Future investigation studying the impact of the proposed weights on both the between-imputation and within-imputation variation may help researchers better understand the inference in the stacked imputation and possibly provide faster tools to estimate the variation.

In this dissertation, we consider the problem under the setting that we have a moderate-dimensional internal data, and each of the external models uses a subset of covariates in the internal study. Although the proposed methods are applicable in high-dimensional settings in theory

and we evaluated them in the settings of multi-dimension X and B, e.g. dim($\mathbf{X}$)=9 and dim($\mathbf{B}$)=5, the performance of applying our methods to higher dimensional data such as the genetic data is unknown. Additional variable selection procedure could be considered if the dimension of $\mathbf{B}$ is large.

APPENDIX A

Appendix of Chapter 2

A.1 Derivation of asymptotic variances for special case 1: Gaussian

A.1.1 Standard MLE:

The log-likelihood is:

$$l = l(\boldsymbol{\gamma}, \sigma_\gamma^2) = \sum_{\mathrm{i}=1}^{\mathrm{n}} logf(\mathrm{Y_i}|\mathrm{X_i}, \mathrm{B_i}; \boldsymbol{\gamma}, \sigma_\gamma^2) = -\frac{\mathrm{n}}{2} log(\sigma_\gamma^2) - \frac{1}{2\sigma_\gamma^2} \sum_{\mathrm{i}=1}^{\mathrm{n}} \left(\mathrm{Y_i} - \gamma_\mathrm{X}\mathrm{X_i} - \gamma_\mathrm{B}\mathrm{B_i}\right)^2$$

Igorning the σ_γ row and column in the information matrix, the expected information matrix for $\hat{\boldsymbol{\gamma}}$ is:

$$\mathbf{I} = -\mathrm{E_{XB}}(\frac{\partial^2 l}{\partial \boldsymbol{\gamma} \boldsymbol{\gamma}^\mathrm{T}}) = \mathrm{n}\frac{\sigma_\mathrm{X}^2}{\sigma_\gamma^2} \begin{pmatrix} 1 & \theta \\ \theta & \theta^2 + \frac{\sigma_\theta^2}{\sigma_\mathrm{X}^2} \end{pmatrix} \tag{A.1}$$

Then we can obtain the asymptotic covariance matrix of $\hat{\boldsymbol{\gamma}}$ by taking the inverse of $\mathbf{I}$:

$$\mathrm{Var}(\hat{\boldsymbol{\gamma}}) = \mathbf{I}^{-1} = \frac{1}{\mathrm{n}} \frac{\sigma_\gamma^2}{\sigma_\theta^2} \begin{pmatrix} \theta^2 + \frac{\sigma_\theta^2}{\sigma_\mathrm{X}^2} & -\theta \\ -\theta & 1 \end{pmatrix}. \tag{A.2}$$

Thus, the asymptotic variance of $\hat{\gamma}_\mathrm{X}$ and $\hat{\gamma}_\mathrm{B}$ are equal to $\frac{\sigma_\gamma^2}{\mathrm{n}\sigma_\theta^2}(\theta^2 + \frac{\sigma_\theta^2}{\sigma_\mathrm{X}^2})$ and $\frac{\sigma_\gamma^2}{\mathrm{n}\sigma_\theta^2}$, respectively.

A.1.2 Approach 1: Synthetic data method

If the synthetic data approach is applied, and under the assumption that the true value of β and σ_β are used to generate the synthetic data, then the combined data will have the same distribution as a dataset of size n+m in which m values of B have been removed. For this particular data structure, it is possible to obtain formulas for the asymptotic variance of the MLE of $\boldsymbol{\gamma}$. In particular, Gourier-

oux and Monfort [1981] gave the exact expression of the MLE and the corresponding asymptotic covariance in such case. The likelihood for the combined data is $\prod_{\mathrm{i=1}}^{\mathrm{n}} f(\mathrm{Y_i}, \mathrm{B_i}|\mathrm{X_i}) \times \prod_{\mathrm{i=n+1}}^{\mathrm{n+m}} f(\mathrm{Y_i}|\mathrm{X_i})$, which can be rewritten as $\prod_{\mathrm{i=1}}^{\mathrm{n+m}} f(\mathrm{Y_i}|\mathrm{X_i}) \times \prod_{\mathrm{i=1}}^{\mathrm{n}} f(\mathrm{B_i}|\mathrm{X_i}, \mathrm{Y_i})$. Based on this likelihood, they introduced a set of transformed parameters, and re-parameterized the distributions 2.5–2.7. They then identified the 1-to-1 relationship among the original parameters and the new set of parameters, which we will explain in the subsequent paragraph.

We obtain the estimators of the original parameters by the re-parameterization method, and then apply the delta method to get the asymptotic variance of $\hat{\gamma}_{\mathrm{B}}$ and $\hat{\gamma}_{\mathrm{X}}$. According to Gourieroux and Monfort [1981], we introduce a set of transformed parameters a, b, c, d, and e, and re-parameterize the distributions 2.5–2.7 as follows:

$$\begin{aligned} \mathrm{Y}|\mathrm{X} &\sim N(b\mathrm{X}, a^2) \\ \mathrm{B}|\mathrm{Y}, \mathrm{X} &\sim N(d\mathrm{Y} + e\mathrm{X}, c^2) \end{aligned}$$

Then we identify the 1-to-1 relationship among the original parameters and the new set of parameters:

$$\begin{aligned} a^2 &= \sigma_\gamma^2 + \gamma_{\mathrm{B}}^2 \sigma_\theta^2 \\ b &= \gamma_{\mathrm{X}} + \theta\gamma_{\mathrm{B}} \\ c^2 &= \frac{\sigma_\gamma^2 \sigma_\theta^2}{a^2} \\ d &= \frac{\gamma_{\mathrm{B}} \sigma_\theta^2}{\mathrm{a}^2} \\ e &= \theta - db \end{aligned} \tag{A.3}$$

The MLE of a, b and their asymptotic variances are easy to obtain from the linear model $\mathrm{Y_i} = b\mathrm{X_i} + u_i, \mathrm{Var}(u_i) = a^2$, where i = 1,...,n+m. Similarly, the MLE of c, d, and e and their asymptotic variances are easy to obtain from the linear model $\mathrm{B_i} = d\mathrm{Y_i} + e\mathrm{X_i} + v_i, \mathrm{Var}(v_i) = c^2$ where i = 1,...,n. The estimators of the original parameters are obtained through the relationship derived from equations (A.3), where

$$\begin{aligned} \theta &= bd + e \\ \sigma_\theta^2 &= a^2 d^2 + c^2 \\ \gamma_{\mathrm{B}} &= \frac{a^2 d}{\sigma_\theta^2} \\ \gamma_{\mathrm{X}} &= \mathrm{b} - \gamma_{\mathrm{B}}\theta \\ \sigma_\gamma^2 &= \frac{a^2 c^2}{\sigma_\theta^2} \end{aligned}$$

and the asymptotic variance of $\hat{\gamma}_{\mathrm{X}}$ and $\hat{\gamma}_{\mathrm{B}}$ can be derived using the delta method:

$$\begin{cases} \mathrm{Var}(\hat{\gamma}_{\mathrm{B}}) = \frac{1}{\mathrm{n}}\big[\frac{\sigma_\gamma^2}{\sigma_\theta^2} + 2(\lambda-1)\frac{\gamma_{\mathrm{B}}^2\sigma_\gamma^4}{\sigma_\beta^4}\big] \\ \mathrm{Var}(\hat{\gamma}_{\mathrm{X}}) = \theta^2\mathrm{Var}(\hat{\gamma}_{\mathrm{B}}) + \frac{1}{\mathrm{n}}\frac{\sigma_\gamma^2}{\sigma_{\mathrm{X}}^2}\frac{\lambda\sigma_\gamma^2+\gamma_{\mathrm{B}}^2\sigma_\theta^2}{\sigma_\beta^2}, \end{cases}$$

Therefore, we find the relative efficiency gain of $\mathrm{Var}(\hat{\boldsymbol{\gamma}}) = \mathrm{Var}(\hat{\gamma}_{\mathrm{X}}, \hat{\gamma}_{\mathrm{B}})^{\mathrm{T}}$ by adding m synthetic data observations compared to the original dataset of size n is

$$\mathrm{ARE}[\mathrm{Var}(\hat{\boldsymbol{\gamma}})] = \mathbf{1} - (1-\lambda)\begin{pmatrix} \frac{2\sigma_\gamma^2\gamma_{\mathrm{B}}^2\sigma_\theta^2}{\sigma_\beta^4} + \frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_{\mathrm{X}}^2\theta^2}\frac{\sigma_\gamma^2(2\sigma_\gamma^2-\sigma_\beta^2)}{\sigma_\beta^4} \\ \frac{2\gamma_{\mathrm{B}}^2\sigma_\theta^2\sigma_\gamma^2}{\sigma_\beta^4} \end{pmatrix},$$

where $\theta = \frac{\beta-\gamma_{\mathrm{X}}}{\gamma_{\mathrm{B}}}$, and $\sigma_\theta^2 = \frac{\sigma_\beta^2-\sigma_\gamma^2}{\gamma_{\mathrm{B}}^2}$. When m gets very large such that $\lambda \approx 0$, ARE[Var($\hat{\gamma_{\mathrm{X}}}$)] = $1 - \frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_{\mathrm{X}}^2\theta^2}\frac{\sigma_\gamma^2}{\sigma_\gamma^2+\gamma_{\mathrm{B}}^2\sigma_\theta^2}\frac{2\sigma_\gamma^2-\sigma_\beta^2}{\sigma_\beta^2} - \frac{2\sigma_\gamma^2\gamma_{\mathrm{B}}^2\sigma_\theta^2}{\sigma_\beta^4}$, and ARE[Var($\hat{\gamma_{\mathrm{B}}}$)] = $1 - \frac{2\sigma_\gamma^2\gamma_{\mathrm{B}}^2\sigma_\theta^2}{\sigma_\beta^4}$. This demonstrates some gain in efficiency for both γ_{X} and γ_{B}.

A.1.3 Approach 2: Constrained MLE

Depending on the information available from the external model Y|X, two possible situations correspond to two different constraints:

- *Approach 2.1: Only the estimated coefficient β is known from model 2.5*

  From model 2.5–2.7, it is easy to see that the constraint is $\theta = \frac{\beta-\gamma_{\mathrm{X}}}{\gamma_{\mathrm{B}}}$, describing the relationship between the unknown variable θ, the known variable β and the target variable $\boldsymbol{\gamma}$. The log-likelihood is given by

$$\begin{aligned} l &= l(\boldsymbol{\gamma}, \theta, \sigma_\gamma^2, \sigma_\theta^2) \\ &= \sum_{\mathrm{i}=1}^{\mathrm{n}} logf(\mathrm{Y_i}, \mathrm{B_i}|\mathrm{X_i}) = \sum_{\mathrm{i}=1}^{\mathrm{n}} logf(\mathrm{Y_i}|\mathrm{X_i}, \mathrm{B_i}; \gamma, \sigma_\gamma^2) + \sum_{\mathrm{i}=1}^{\mathrm{n}} logf(\mathrm{B_i}|\mathrm{X_i}; \theta, \sigma_\theta^2) \\ &= -\frac{\mathrm{n}}{2}log(\sigma_\gamma^2) - \frac{1}{2\sigma_\gamma^2}\sum_{\mathrm{i}=1}^{\mathrm{n}}\big(\mathrm{Y_i} - \gamma_{\mathrm{X}}\mathrm{X_i} - \gamma_{\mathrm{B}}\mathrm{B_i}\big)^2 - \frac{\mathrm{n}}{2}log(\sigma_\theta^2) - \frac{1}{2\sigma_\theta^2}\sum_{\mathrm{i}=1}^{\mathrm{n}}\big(\mathrm{B_i} - \theta\mathrm{Xi}\big)^2 \end{aligned} \tag{A.4}$$

  The goal is to maximize the log-likelihood A.4 over $\boldsymbol{\gamma}$, σ_γ and σ_θ subject to the constraint $\theta = \theta^* = \frac{\beta^*-\gamma_{\mathrm{X}}}{\gamma_{\mathrm{B}}}$. By replacing θ with θ^*, taking the second derivative over $\boldsymbol{\gamma}$, and taking the inverse of the information matrix, we obtain the asymptotic variance of $\hat{\boldsymbol{\gamma}}$:

$$\mathrm{Var}(\hat{\boldsymbol{\gamma}}) = \mathbf{I}^{-1} = \frac{1}{\mathrm{n}}\frac{\sigma_\gamma^2}{\sigma_\theta^2}\begin{pmatrix} \theta^{*2} + \frac{\sigma_\theta^4\gamma_{\mathrm{B}}^2}{\sigma_\gamma^2+\sigma_\theta^2\gamma_{\mathrm{B}}^2}\sigma_{\mathrm{X}}^{-2} & -\theta^* \\ -\theta^* & 1 \end{pmatrix}. \tag{A.5}$$

Thus, the ARE of $\mathrm{Var}(\hat{\boldsymbol{\gamma}})$ from the constrained MLE compared to the standard MLE is

$$\mathrm{ARE}[\mathrm{Var}(\hat{\boldsymbol{\gamma}})] = \begin{pmatrix} 1 - \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_{\mathrm{X}}^2 \theta^{*2}} \frac{\sigma_\gamma^2}{\sigma_\gamma^2 + \gamma_{\mathrm{B}}^2 \sigma_\theta^2} \\ 1 \end{pmatrix},$$

where we notice that there is some gain in efficiency for γ_{X} but no gain in efficiency for γ_{B}. We can see that the largest gain in efficiency is when γ_{B}, θ and σ_{X} are small.

- *Approach 2.2: Both of the estimated coefficient β and the standard deviation σ_β are known from model 2.5*

  In this situation, knowing the true σ_β gives us more information which is incorporated through an additional constraint. In addition to the constraint $\theta = \theta^* = \frac{\beta - \gamma_{\mathrm{X}}}{\gamma_{\mathrm{B}}^*}$ derived in approach 2.1, we add another constraint $\mathrm{Var}(\mathrm{Y}|\mathrm{X}) = \sigma_\beta^2 = \gamma_{\mathrm{B}}^2 \sigma_\theta^2 + \sigma_\gamma^2$, where $\sigma_\theta^2 = \sigma_\theta^{*2} = \frac{\sigma_\beta^{*2} - \sigma_\gamma^2}{\gamma_{\mathrm{B}}^2}$. Then we maximize log-likelihood A.4 with respect to $\boldsymbol{\gamma}$ and σ_γ^2 at fixed values $\sigma_\theta^2 = \sigma_\theta^{*2}, \theta = \theta^*$. Note that different from approach 2.1, σ_γ^2 and $\boldsymbol{\gamma}$ are not independent anymore. Thus, we need to consider σ_γ^2 in the information matrix, and take the inverse of a 3×3 matrix to get the correct asymptotic variance. Let $\boldsymbol{\phi} = (\boldsymbol{\gamma}, \sigma_\gamma^2)^{\mathrm{T}}$,

$$\mathbf{I} = -\mathrm{E}_{\mathrm{XB}}\left(\frac{\partial^2 l}{\partial \boldsymbol{\phi} \boldsymbol{\phi}^{\mathrm{T}}}\right) = \mathrm{n} \begin{pmatrix} (\frac{1}{\sigma_\gamma^2} + \frac{1}{\gamma_{\mathrm{B}}^2 \sigma_\theta^{*2}})\sigma_{\mathrm{X}}^2 & (\frac{1}{\sigma_\gamma^2} + \frac{1}{\gamma_{\mathrm{B}}^2 \sigma_\theta^{*2}})\sigma_{\mathrm{X}}^2 \theta^* & 0 \\ (\frac{1}{\sigma_\gamma^2} + \frac{1}{\gamma_{\mathrm{B}}^2 \sigma_\theta^{*2}})\sigma_{\mathrm{X}}^2 \theta^* & (\frac{1}{\sigma_\gamma^2} + \frac{1}{\gamma_{\mathrm{B}}^2 \sigma_\theta^{*2}})(\sigma_\theta^{*2} + \sigma_{\mathrm{X}}^2 \theta^{*2}) + \frac{1}{\gamma_{\mathrm{B}}^2} & \frac{1}{\sigma_\theta^{*2} \gamma_{\mathrm{B}}^3} \\ 0 & \frac{1}{\sigma_\theta^{*2} \gamma_{\mathrm{B}}^3} & \frac{1}{2}(\frac{1}{\sigma_\gamma^4} + \frac{1}{\gamma_{\mathrm{B}}^4 \sigma_\theta^{*4}}) \end{pmatrix},$$

  By taking the inverse of $\mathbf{I}$, we can get the asymptotic variance of $\hat{\boldsymbol{\gamma}}$:

$$\begin{cases} \mathrm{Var}(\hat{\gamma}_{\mathrm{B}}) & = \frac{1}{\mathrm{n}} \frac{\sigma_\gamma^2}{\sigma_\theta^{*2}} \frac{\sigma_\gamma^4 + \gamma_{\mathrm{B}}^4 \sigma_\theta^{*4}}{(\sigma_\gamma^2 + \gamma_{\mathrm{B}}^2 \sigma_\theta^{*2})^2} = \frac{1}{\mathrm{n}} \frac{\gamma_{\mathrm{B}}^2 \sigma_\gamma^2}{\sigma_\beta^{*2} - \sigma_\gamma^2} \frac{\sigma_\gamma^4 + (\sigma_\beta^{*2} - \sigma_\gamma^2)^2}{\sigma_\beta^{*4}} \\ \mathrm{Var}(\hat{\gamma}_{\mathrm{X}}) & = \frac{1}{\mathrm{n}} \frac{\sigma_\gamma^2}{\sigma_\theta^{*2}} \frac{1}{(\sigma_\gamma^2 + \gamma_{\mathrm{B}}^2 \sigma_\theta^{*2})^2} \left[(\sigma_\theta^{*2} \sigma_{\mathrm{X}}^{-2} + \theta^{*2})(\sigma_\gamma^4 + \gamma_{\mathrm{B}}^4 \sigma_\theta^{*4}) - (\sigma_\gamma^2 - \gamma_{\mathrm{B}}^2 \sigma_\theta^{*2}) \sigma_\gamma^2 \sigma_\theta^{*2} \sigma_{\mathrm{X}}^{-2} \right] \\ & = (\sigma_\theta^{*2} \sigma_{\mathrm{X}}^{-2} + \theta^{*2}) \mathrm{Var}(\hat{\gamma}_{\mathrm{B}}) - \frac{1}{\mathrm{n}} \sigma_\gamma^4 \sigma_{\mathrm{X}}^{-2} \frac{\sigma_\gamma^2 - \gamma_{\mathrm{B}}^2 \sigma_\theta^{*2}}{\sigma_\beta^{*4}} \end{cases}$$

  Thus, we can obtain the identical ARE to the synthetic data method (approach 1 in Appendix A.1.2). This demonstrates the asymptotic equivalence of the synthetic data approach with large m compared to the constrained ML approach that uses knowledge of all the parameters in the $\mathrm{Y}|\mathrm{X}$ distribution.

A.1.4 Approach 3: CSPML

This approach assumes that β is known, but does not assume that σ_β is known. To calculate the asymptotic variance of $\hat{\boldsymbol{\gamma}}$ in this approach, we need three matrices $\mathbf{I}$, $\mathbf{C}$ and $\mathbf{L}$. After some algebra,

it can be shown that $\mathbf{C} = \frac{\sigma_{\mathrm{X}}^2}{\sigma_\beta^2}(1, \theta^*)^{\mathrm{T}}, \mathrm{L} = \frac{\mathrm{n}\gamma_{\mathrm{B}}^2\sigma_\theta^*\sigma_{\mathrm{X}}^2}{\sigma_\beta^{*4}}$. Thus,

$$\mathrm{Cov}(\hat{\boldsymbol{\gamma}}) = (\mathbf{I} + \mathbf{C}\mathbf{L}^{-1}\mathbf{C}^{\mathrm{T}})^{-1} = \frac{1}{\mathrm{n}}\frac{\sigma_\gamma^2}{\sigma_\theta^2}\begin{pmatrix} \theta^{*2} + \frac{\sigma_\theta^4\gamma_{\mathrm{B}}^2}{\sigma_\gamma^2+\sigma_\theta^2\gamma_{\mathrm{B}}^2}\sigma_{\mathrm{X}}^{-2} & -\theta^* \\ -\theta^* & 1 \end{pmatrix},$$

which is identical to approach 2.1 in Appendix A.1.3.

A.2 Derivation of asymptotic variances for special case 2: binary

A.2.1 Standard MLE

We will be using the following notation: $\mathrm{S} \equiv \mathrm{S}_\gamma(\mathrm{X}, \mathrm{B}) = \gamma_0 + \gamma_{\mathrm{X}}\mathrm{X} + \gamma_{\mathrm{B}}\mathrm{B} + \gamma_{\mathrm{XB}}\mathrm{XB}$, $\mathrm{M} \equiv \mathrm{M}_\beta(\mathrm{X}) = \beta_0 + \beta_1\mathrm{X}$ and $\mathrm{K} \equiv \mathrm{K}_\theta(\mathrm{X}) = \theta_0 + \theta_1\mathrm{X}$. The ML estimators are the solution of maximizing $\prod_{\mathrm{i}=1}^{\mathrm{n}} f(\mathrm{Y_i}|\mathrm{X_i}, \mathrm{B_i}; \boldsymbol{\gamma})$, i.e. $\max_{\boldsymbol{\gamma}}\Big\{\sum_{\mathrm{i}=1}^{\mathrm{n}}[\mathrm{Y_i}\mathrm{S_i} - log\{1 + exp(\mathrm{S_i})\}]\Big\}$.

In the tri-binary case, since $\mathrm{X} = \mathrm{X}^2$, and $\mathrm{B} = \mathrm{B}^2$, the Fisher information matrix has the following form:

$$\mathbf{I} = \mathrm{E_{XB}I(X, B)} = \mathrm{E_{XB}}\Big\{expit(\mathrm{S})[1 - expit(\mathrm{S})]\begin{pmatrix} 1 & \mathrm{X} & \mathrm{B} & \mathrm{XB} \\ \mathrm{X} & \mathrm{X} & \mathrm{XB} & \mathrm{XB} \\ \mathrm{B} & \mathrm{XB} & \mathrm{B} & \mathrm{XB} \\ \mathrm{XB} & \mathrm{XB} & \mathrm{XB} & \mathrm{XB} \end{pmatrix}\Big\}.$$

There are four possible combinations of binary (X, B), i.e. (0,0), (0,1), (1,0) and (1,1). Thus, the expectation terms of the matrix $\mathrm{I}(\mathrm{X}, \mathrm{B})$ can be obtained through $\frac{1}{\mathrm{n}}\sum_{\mathrm{a,b}\in\{0,1\}}\mathrm{I(a, b)P(X = a, B = b)}$. The asymptotic variance of $\hat{\boldsymbol{\gamma}}$ is given by:

$$\begin{cases} \mathrm{Var}(\hat{\gamma}_0) = \frac{1}{\mathrm{n}}(\frac{1}{\mathrm{P(BXY=001)}} + \frac{1}{\mathrm{P(BXY=000)}}), \\ \mathrm{Var}(\hat{\gamma}_{\mathrm{X}}) = \frac{1}{\mathrm{n}}\sum_{\mathrm{a,b}\in\{0,1\}}\frac{1}{\mathrm{P(BXY=0ab)}}, \\ \mathrm{Var}(\hat{\gamma}_{\mathrm{B}}) = \frac{1}{\mathrm{n}}(\frac{1}{\mathrm{P(B=1|XY=01)P(BXY=001)}} + \frac{1}{\mathrm{P(B=1|XY=00)P(BXY=000)}}), \\ \mathrm{Var}(\hat{\gamma}_{\mathrm{XB}}) = \frac{1}{\mathrm{n}}\sum_{\mathrm{a,b}\in\{0,1\}}\frac{1}{\mathrm{P(B=1|XY=ab)P(BXY=0ab)}}, \end{cases}$$

where $\mathrm{P(BXY = 0ab)}$ is the probability of the (B=0, X=a, Y=b) combination, and $\mathrm{P(B = 1|XY = ab)}$ is the probability of B=1 given X=a and Y=b, a,b$\in \{0, 1\}$.

A.2.2 Approach 1: Synthetic data method

Motivated by the MLE in the missing data problem (Little, 1992), we re-formate our target likelihood as follows:

$$\begin{aligned}\prod_{i=1}^{n} f(Y_i, X_i, B_i) \prod_{i=n+1}^{m+n} f(X_i, Y_i) &= \prod_{i=1}^{n} f(X_i, Y_i) f(B_i|X_i Y_i) \prod_{i=n+1}^{m+n} f(X_i, Y_i) \\ &= \prod_{i=1}^{n} f(B_i|X_i, Y_i) \prod_{i=1}^{m+n} f(X_i, Y_i),\end{aligned} \tag{A.6}$$

where $f(B_i|X_i, Y_i)$ and $f(X_i, Y_i)$ are independent from each other. The goal is to maximize likelihood A.6 over $\boldsymbol{\gamma}$.

Let $P(XY = ab) \equiv Pr(X_i = a, Y_i = b)$, a, b $\in \{0, 1\}$, i = 1,...,m+n. With the constraint $\sum_{a,b\in\{0,1\}} P(XY = ab) = 1$, there are three unknown variables in $\prod_{i=1}^{m+n} f(X_i, Y_i)$, i.e. P(XY=ab), $a, b \in \{0, 1\}$. Denote $P(B = 1|XY = ab) \equiv Pr(B_i = 1|X_i = a, Y_i = b)$, i = 1,...,n. Similarly, since there are four different combination of a and b, there are four unknown parameters in $\prod_{i=1}^{n} f(B_i|X_i, Y_i)$, i.e. $P(B = 1|XY = ab)$, each of which is independent of others. By plugging in the four possible combinations of X and B into model 2.8 in Chapter 2, we can easily derive the expressions for $\boldsymbol{\gamma}$ as presented in Table A.1.

Table A.1: Formulas of $\boldsymbol{\gamma}$ in terms of $P(B|XY)$ and $P(XY)$

(X, B) combination	Transformation of model 2.8
(0, 0)	$\gamma_0 = \log\frac{P(B=0\|XY=01)P(XY=01)}{P(B=0\|XY=00)P(XY=00)}$
(1, 0)	$\gamma_0 + \gamma_X = \log\frac{P(B=0\|XY=11)P(XY=11)}{P(B=0\|XY=10)P(XY=10)}$
(0, 1)	$\gamma_0 + \gamma_B = \log\frac{P(B=1\|XY=01)P(XY=01)}{P(B=1\|XY=00)P(XY=00)}$
(1, 1)	$\gamma_0 + \gamma_X + \gamma_B + \gamma_{XB} = \log\frac{P(B=1\|XY=11)P(XY=11)}{P(B=1\|XY=10)P(XY=10)}$

Let $mn(a, b)$ denote the number of observations with (X = a,Y = b) in the sample size of m+n. Since $mn(a, b) \sim Multinomial(m + n, P(XY = ab))$, we can easily obtain the MLE of $P(XY)$, and the corresponding estimated covariance as follows:

$$\begin{cases}\hat{P}(XY = ab) = \frac{mn(a,b)}{m+n} \\ Var[\hat{P}(XY = ab)] = \hat{P}(XY = ab)[1 - \hat{P}(XY = ab)] \\ Cov[\hat{P}(XY = ab), \hat{P}(XY = a'b')] = -\hat{P}(XY = ab)\hat{P}(XY = a'b')\end{cases}$$

Let n(a,b) be the number of observations with $(\mathrm{X} = \mathrm{a}, \mathrm{Y} = \mathrm{b})$ in the sample of size n, and $\mathrm{n}(\mathrm{B} = 1|\mathrm{XY} = \mathrm{ab})$ be the count of B = 1 given (X = a, Y = b). Since $\mathrm{n}(\mathrm{B} = 1|\mathrm{XY} = \mathrm{ab}) \sim \mathrm{Bin}(\mathrm{n}(\mathrm{a}, \mathrm{b}), \mathrm{P}(\mathrm{B} = 1|\mathrm{XY} = \mathrm{ab}))$, the MLE of $\mathrm{P}(\mathrm{B}|\mathrm{X}, \mathrm{Y})$ and its estimated covariance can be expressed as:

$$\begin{cases} \hat{\mathrm{P}}(\mathrm{B} = 1|\mathrm{XY} = \mathrm{ab}) = \frac{\mathrm{n(B=1|XY=ab)}}{\mathrm{n(a,b)}} \\ \mathrm{Var}(\hat{\mathrm{P}}(\mathrm{B} = 1|\mathrm{XY} = \mathrm{ab})) = \frac{\hat{\mathrm{P}}(\mathrm{B=1|XY=ab})\hat{\mathrm{P}}(\mathrm{B=0|XY=ab})}{\mathrm{n(a,b)}} \\ \mathrm{Cov}(\hat{\mathrm{P}}(\mathrm{B} = 1|\mathrm{XY} = \mathrm{ab}), \hat{\mathrm{P}}(\mathrm{B} = 1|\mathrm{XY} = \mathrm{a'b'})) = 0 \end{cases}$$

Therefore, the MLE of $\boldsymbol{\gamma}$ can be expressed as:

$$\begin{cases} \hat{\gamma}_0 = \log \frac{\hat{\mathrm{P}}(\mathrm{B=0|XY=01})}{\hat{\mathrm{P}}(\mathrm{B=0|XY=00})} \frac{\hat{\mathrm{P}}(\mathrm{XY=01})}{\hat{\mathrm{P}}(\mathrm{XY=00})} \\ \hat{\gamma}_{\mathrm{X}} = \log \frac{\hat{\mathrm{P}}(\mathrm{B=0|XY=11})\hat{\mathrm{P}}(\mathrm{B=0|XY=00})}{\hat{\mathrm{P}}(\mathrm{B=0|XY=10})\hat{\mathrm{P}}(\mathrm{B=0|XY=01})} \frac{\hat{\mathrm{P}}(\mathrm{XY=11})\hat{\mathrm{P}}(\mathrm{XY=00})}{\hat{\mathrm{P}}(\mathrm{XY=10})\hat{\mathrm{P}}(\mathrm{XY=01})} \\ \hat{\gamma}_{\mathrm{B}} = \log \frac{\hat{\mathrm{P}}(\mathrm{B=1|XY=01})}{\hat{\mathrm{P}}(\mathrm{B=0|XY=01}} \frac{\hat{\mathrm{P}}(\mathrm{B=0|XY=00})}{\hat{\mathrm{P}}(\mathrm{B=1|XY=00})} \\ \hat{\gamma}_{\mathrm{XB}} = \log \frac{\hat{\mathrm{P}}(\mathrm{B=1|XY=11})\hat{\mathrm{P}}(\mathrm{B=1|XY=00})}{\hat{\mathrm{P}}(\mathrm{B=1|XY=10})\mathrm{P}(\mathrm{B=1|XY=01})} \frac{\hat{\mathrm{P}}(\mathrm{B=0|XY=10})\hat{\mathrm{P}}(\mathrm{B=0|XY=01})}{\hat{\mathrm{P}}(\mathrm{B=0|XY=00})\hat{\mathrm{P}}(\mathrm{B=0|XY=11})} \end{cases}$$

According to the delta method and replacing the estimated proportions by the corresponding probabilities, we obtain the asymptotic variances as follow:

$$\begin{cases} \mathrm{Var}(\hat{\gamma}_0) = \frac{1}{\mathrm{n}} \sum\limits_{\mathrm{(a,b)}\in\{\mathrm{(0,1),(0,0)}\}} \frac{\mathrm{P(B=1|XY=ab)}}{\mathrm{P(BXY=0ab)}} + \frac{1}{\mathrm{m+n}} \sum\limits_{\mathrm{(a,b)}\in\{\mathrm{(0,1),(0,0)}\}} \frac{1}{\mathrm{P(XY=ab)}} \\ \mathrm{Var}(\hat{\gamma}_{\mathrm{X}}) = \frac{1}{\mathrm{n}} \sum\limits_{\mathrm{a,b}\in\{0,1\}} \frac{\mathrm{P(B=1|XY=ab)}}{\mathrm{P(BXY=0ab)}} + \frac{1}{\mathrm{m+n}} \sum\limits_{\mathrm{a,b}\in\{0,1\}} \frac{1}{\mathrm{P(XY=ab)}} \\ \mathrm{Var}(\hat{\gamma}_{\mathrm{B}}) = \frac{1}{\mathrm{n}} \sum\limits_{\mathrm{(a,b)}\in\{\mathrm{(0,1),(0,0)}\}} \frac{1}{\mathrm{P(B=1|XY=ab)P(BXY=0ab)}} \\ \mathrm{Var}(\hat{\gamma}_{\mathrm{XB}}) = \frac{1}{\mathrm{n}} \sum\limits_{\mathrm{a,b}\in\{0,1\}} \frac{1}{\mathrm{P(B=1|XY=ab)P(BXY=0ab)}} \end{cases}$$

where $\mathrm{P}(\mathrm{BXY} = 0\mathrm{ab}) = \mathrm{P}(\mathrm{B} = 0|\mathrm{XY} = \mathrm{ab})\mathrm{P}(\mathrm{XY} = \mathrm{ab})$.

Therefore, we find that the ARE of $\mathrm{Var}(\hat{\boldsymbol{\gamma}})$ by adding m synthetic data observations compared to the original dataset of size n is

$$\mathrm{ARE}[\mathrm{Var}(\hat{\boldsymbol{\gamma}})] = \mathbf{1} - (1 - \lambda) \begin{pmatrix} \frac{\sum\limits_{\mathrm{(a,b)}\in\{\mathrm{(0,1)(0,0)}\}} 1/\mathrm{P(XY=ab)}}{\sum\limits_{\mathrm{(a,b)}\in\{\mathrm{(0,1)(0,0)}\}} 1/\mathrm{P(BXY=0ab)}} \\ \frac{\sum\limits_{\mathrm{a,b}\in\{0,1\}} 1/\mathrm{P(XY=ab)}}{\sum\limits_{\mathrm{a,b}\in\{0,1\}} 1/\mathrm{P(BXY=0ab)}} \\ 0 \\ 0 \end{pmatrix} \tag{A.7}$$

A.2.3 Approach 2: Constrained MLE

When the summary-level information from model 2.9 is available in the form of coefficient estimates $\boldsymbol{\beta}$, the constrained MLE is the solution of maximizing $\prod_{\mathrm{i}=1}^{\mathrm{n}} f(\mathrm{Y_i}|\mathrm{X_i},\mathrm{B_i};\boldsymbol{\gamma}) f(\mathrm{B_i}|\mathrm{X_i};\boldsymbol{\theta})$ subject to the constraint that $\Pr(\mathrm{Y}=1|\mathrm{X}=\mathrm{x};\boldsymbol{\beta}) = \sum_{\mathrm{b}=0}^{1} \Pr(\mathrm{Y}=1|\mathrm{X}=\mathrm{x},\mathrm{B}=\mathrm{b};\boldsymbol{\gamma})\Pr(\mathrm{B}=\mathrm{b}|\mathrm{X}=\mathrm{x};\boldsymbol{\theta})$. The log-likelihood can be rewritten as

$$\max_{\boldsymbol{\gamma}} \Big\{ \sum_{\mathrm{i}=1}^{\mathrm{n}} [\mathrm{Y_i S_i} - log\{1+exp(\mathrm{S_i})\} + \mathrm{B_i K_i} - log\{1+exp(\mathrm{K_i})\}] \Big\}$$

From the constraints, we can write $\boldsymbol{\theta}$ as a function of $\boldsymbol{\gamma}$ in the following way:

$$\begin{cases} \theta_0(\boldsymbol{\gamma}) & = logit\{\frac{expit(\beta_0)-expit(\gamma_0)}{expit(\gamma_0+\gamma_\mathrm{B})-expit(\gamma_0)}\} \\ \theta_1(\boldsymbol{\gamma}) & = logit\{\frac{expit(\beta_0+\beta_1)-expit(\gamma_0+\gamma_\mathrm{X})}{expit(\gamma_0+\gamma_\mathrm{X}+\gamma_\mathrm{B}+\gamma_\mathrm{XB})-expit(\gamma_0+\gamma_\mathrm{X})}\} - \theta_0(\boldsymbol{\gamma}) \end{cases}$$

Then K becomes $\mathrm{K}_\gamma(\mathrm{X}) = \theta_0(\boldsymbol{\gamma}) + \theta_1(\boldsymbol{\gamma})\mathrm{X}$. Denote $\begin{cases} \sigma_{\beta_{0\mathrm{j}}} & \equiv \frac{\partial}{\partial\gamma_\mathrm{j}}\theta_0(\boldsymbol{\gamma}) \\ \sigma_{\beta_{1\mathrm{j}}} & \equiv \frac{\partial}{\partial\gamma_\mathrm{j}}\theta_1(\boldsymbol{\gamma}) \end{cases}$ where j = 0, 1, 2, 3. The asymptotic variance of $\hat{\boldsymbol{\gamma}}$ can be derived through the 4×4 matrix $\frac{1}{\mathrm{n}}\{\mathrm{E_{XB}}[\mathrm{E_{Y|XB}}(\mathbf{u_\gamma}\mathbf{u_\gamma}^\mathrm{T})]\}^{-1}$, where

$$\begin{aligned} \mathbf{u_\gamma} &= \frac{\partial}{\partial\boldsymbol{\gamma}} log\{f(\mathrm{Y},\mathrm{B}|\mathrm{X};\boldsymbol{\gamma},\boldsymbol{\theta}(\boldsymbol{\gamma}))\} \\ &= \begin{pmatrix} (\sigma_{\beta_{00}}+\sigma_{\beta_{10}}\mathrm{X})(\mathrm{B}-expit(\mathrm{K}))+\mathrm{Y}-expit(\mathrm{S}) \\ (\sigma_{\beta_{01}}+\sigma_{\beta_{11}}\mathrm{X})(\mathrm{B}-expit(\mathrm{K})))+(\mathrm{Y}-expit(\mathrm{S}))\mathrm{X} \\ (\sigma_{\beta_{02}}+\sigma_{\beta_{12}}\mathrm{X})(\mathrm{B}-expit(\mathrm{K})))+(\mathrm{Y}-expit(\mathrm{S}))\mathrm{B} \\ (\sigma_{\beta_{03}}+\sigma_{\beta_{13}}\mathrm{X})(\mathrm{B}-expit(\mathrm{K})))+(\mathrm{Y}-expit(\mathrm{S}))\mathrm{XB} \end{pmatrix} \end{aligned}$$

Since all Y,X and B are binary variables, there are eight possible combinations of $(\mathrm{Y},\mathrm{X},\mathrm{B})$. Thus, the expectation term in the matrix $\mathrm{E}(\mathbf{u_\gamma}\mathbf{u_\gamma}^\mathrm{T})$ can be obtained through

$$\frac{1}{\mathrm{n}} \sum_{\mathrm{a,b,c}\in\{0,1\}} \mathrm{u_\gamma u_\gamma^T P}(\mathrm{Y}=\mathrm{a},\mathrm{X}=\mathrm{b},\mathrm{B}=\mathrm{c}).$$

A variation of the above approach is when the external summary information comes in the form of the predicted probability, for any given X, i.e. we are simply provided with $\bar{\mathrm{P}}(\mathrm{X_i}) = \hat{\Pr}(\mathrm{Y_i}=1|\mathrm{X_i})$. In such case, it is easy to construct an estimation method that uses the estimated probability as a constraint. In addition, in the special case being considered here where Y and X are binary, it is easy to see that knowing $\bar{\mathrm{P}}(0)$ and $\bar{\mathrm{P}}(1)$ is equivalent to knowing $\beta_0 = logit[\bar{\mathrm{P}}(0)]$ and $\beta_1 = logit[\bar{\mathrm{P}}(1)] - logit[\bar{\mathrm{P}}(0)]$, so this also fits into the above framework to obtain the asymptotic

variance of $\hat{\boldsymbol{\gamma}}$.

A.2.4 Approach 3: CSPML

By implementing the specific distribution into the given formulas for $\mathbf{I}$, $\mathbf{C}$ and $\mathbf{L}$, we find that $\mathbf{I}$ is the same as the information matrix in approach 1 in Appendix A.2.2, where $\mathbf{C}_{4\times 2}$ is the first two columns of matrix $\mathbf{I}_{4\times 4}$, and

$$\mathbf{L} = \mathrm{E}_{\mathrm{XB}}\Big\{ \mathit{expit}(\mathrm{S})[1-\mathit{expit}(\mathrm{M})] - \mathit{expit}(\mathrm{M})[1-\mathit{expit}(\mathrm{S})] \begin{pmatrix} 1 & \mathrm{X} \\ \mathrm{X} & \mathrm{X}^2 \end{pmatrix} \Big\},$$

where $\mathrm{S} \equiv \mathrm{S}_{\gamma}(\mathrm{X},\mathrm{B}) = \gamma_0 + \gamma_{\mathrm{X}}\mathrm{X} + \gamma_{\mathrm{B}}\mathrm{B} + \gamma_{\mathrm{XB}}\mathrm{XB}$ and $\mathrm{M} \equiv \mathrm{M}_{\beta}(\mathrm{X}) = \beta_0 + \beta_1\mathrm{X}$. The calculation of $\mathbf{L}$ is simple under the situation where X and B are both binary. Then $\mathbf{I}$, $\mathbf{C}$ and $\mathbf{L}$ can be combined to give the variance of $\hat{\boldsymbol{\gamma}}$.

Although we do not write out the formulas of ARE[Var($\hat{\boldsymbol{\gamma}}$)] for approaches 2 and 3 in Appendix A.2.3 and A.2.4, respectively, they are numerically identical to the equation A.7 when $\lambda = 0$.

APPENDIX B

Appendix of Chapter 3

B.1 Summary of Features and Assumptions

Figure B.1 summarizes the key features and assumptions required in the internal-data-only regression, CSPML, EB, and the proposed approaches.

In the CSPML approach, the estimated $\mathrm{Y}|\mathrm{X}$ model from the external data is assumed to be the best fitted model in the class, but that class of models may not include the true model. A common example is that when the true $\mathrm{Y}|\mathrm{X}, \mathrm{B}$ model belongs to the generalized linear model (GLM) family, it does not necessarily imply the $\mathrm{Y}|\mathrm{X}$ model belongs to the same class of GLMs (e.g. when $\mathrm{Y}|\mathrm{X}, \mathrm{B}$ is linear model, $\mathrm{Y}|\mathrm{X}$ could still be linear model if $\mathrm{B}|\mathrm{X}$ is linear; but when $\mathrm{Y}|\mathrm{X}, \mathrm{B}$ is logistic model, it is common to fit a logistic model for $\mathrm{Y}|\mathrm{X}$, but the truth is not a logistic model as collapsibility does not hold for the logit link).

As showed in Figure B.1, the CSPML approach implicitly requires the joint distribution of (Y, X, B) to be identical in the internal and the external population, so it requires more than the mean exchangeability assumption, i.e. it requires more than the conditional distribution of $\mathrm{Y}|\mathrm{X}, \mathrm{B}$ to be the same. The CSPML will essentially always be biased if the $\mathrm{Y}|\mathrm{X}, \mathrm{B}$ distributions differ. As demonstrated in simulations, the CSPML can also be biased if the two populations only differ in the (X,B) distribution. The bias can be small or it can be large depending on the details of the situation. The EB estimator ameliorates differences between the population distributions by down-weighting the CSPML estimator if the data suggests it is biased.

Because the EB will down-weight the CSPML estimator if the lack of transportability leads to a poor CSPML estimate, it is robust to departures from the assumption of full transportability in specific external populations. The EB estimator is a shrinkage estimator that posits an additional stochastic framework for the underlying true parameter $\boldsymbol{\gamma} \sim \mathrm{N}(\boldsymbol{\gamma}_0, \mathbf{A})$. The stochastic framework connects the internal estimator $\hat{\boldsymbol{\gamma}}_{\mathrm{I}}$ and the CSPML estimator $\hat{\boldsymbol{\gamma}}_{\mathrm{CSPML}}$, and the difference $\hat{\boldsymbol{\gamma}}_{\mathrm{I}} - \hat{\boldsymbol{\gamma}}_{\mathrm{CSPML}}$ is a measure of the distributional similarity of the joint distribution (Y, X, B) between the internal and the external population.

Approach*		Estimator	Target (Y|X, B; γ)		Additional Assumptions	Bias of $\hat{\gamma}$	Precision gains in estimating X coefficients
			Internal Study	External Study			
Internal Data Only		$\hat{\gamma}_I$	Correctly specified	-	-	Asymptotically unbiased	-
CSPML		$\hat{\gamma}_{CSPML}$	Correctly specified	Same as the internal study	(X, B) is the same in the internal and external studies ($Y\|X; \hat{\beta}$) is the best fitted model in the class, but the class of models may not include the true model	Potential for bias when the joint distribution of (Y, X, B) is different in the two populations, which could arise from various forms of misspecification. E.g. B|X, Y|X,B, Y|X or B.	The variance of X coefficients will usually decrease even when the joint distribution of (Y, X, B) is different in the two populations
EB		$\hat{\gamma}_{EB}$, a shrinkage estimator that can be viewed as a weighted average of $\hat{\gamma}_I$ and $\hat{\gamma}_{CSPML}$	Correctly specified	-	$\gamma \sim N(\gamma_0, A)$, relating $\hat{\gamma}_I$ and $\hat{\gamma}_{CSPML}$; The difference $\hat{\gamma}_I - \hat{\gamma}_{CSPML}$ is used to measure the distributional similarity of the joint distribution (Y, X, B) between the internal and the external population.	Asymptotically unbiased Small to moderate bias in finite samples. Applies less weight to $\hat{\gamma}_{CSPML}$ if the lack of transportability leads to a poor $\hat{\gamma}_{CSPML}$.	The smaller the difference between $\hat{\gamma}_I$ and $\hat{\gamma}_{CSPML}$, the larger the efficiency gain. Efficiency gains diminish as the difference of $\hat{\gamma}_I$ and $\hat{\gamma}_{CSPML}$ becomes large.
Proposed	**OCWE**	$\hat{\gamma}_{OCWE}$	Same as the EB approach			A weighted average of $\hat{\gamma}_{EB}$'s, where the same weight is applied to each coefficient within $\hat{\gamma}_{EB}$	
	SC-Learner	$\hat{\gamma}_{SC-Learner}$				A weighted average of $\hat{\gamma}_{EB}$'s, where a different weight is attached to each coefficient.	

*All approaches assume the parametric settings: g[E(Y|X, B)] is linear in (X, B); and g[E(Y|X)] is linear in X, where g(.) is the known link function.
Abbreviations: CSPML=Constrained Semi-Parametric Maximum Likelihood; EB=Empirical Bayes; γ_0=the underlying true parameter of γ; A=some covariance matrix of γ; E=optimal covariace weighted estimator; SC-Learner=selective coefficient learner.

Figure B.1: Summary of features and assumptions required in the direct regression, CSPML, EB and the proposed approaches.

B.2 Derivation of the asymptotic Distribution

Let $S_i(\gamma)$ denote the score function of $Y|\mathbf{X}, B$ distribution, and u, B, C and L defined in equation 3.2 from Section 3.2.4. The MLE of internal data solves $\frac{1}{n}\sum_{i=1}^{n} S_i(\gamma_I) = 0$ to obtain $\hat{\gamma}_I$. CSPML solves $\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} S_i(\gamma_{CSPML}) + \frac{\frac{\partial u(\gamma_{CSPML})}{\partial \gamma_{CSPML}}}{1-\lambda^T u(\gamma_{CSPML})}\lambda \\ \frac{u(\gamma_{CSPML})}{1-\lambda^T u(\gamma_{CSPML})} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ to obtain $\hat{\gamma}_{CSPML}$ and $\hat{\lambda}$. We obtain the asymptotic distribution of target parameter $\boldsymbol{\eta} = (\gamma_{CSPML}, \lambda, \gamma_I)^T$ by the following steps:

1. Re-write the estimating equation $\mathrm{g}(\boldsymbol{\eta}) = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \begin{pmatrix} \mathrm{S_i}(\boldsymbol{\gamma}_{\mathrm{CSPML}}) + \frac{\frac{\partial \mathrm{u}(\boldsymbol{\gamma}_{\mathrm{CSPML}})}{\partial \boldsymbol{\gamma}_{\mathrm{CSPML}}}}{1-\boldsymbol{\lambda}^{\mathrm{T}}\mathrm{u}(\boldsymbol{\gamma}_{\mathrm{CSPML}})}\boldsymbol{\lambda} \\ \frac{\mathrm{u}(\boldsymbol{\gamma}_{\mathrm{CSPML}})}{1-\boldsymbol{\lambda}^{\mathrm{T}}\mathrm{u}(\boldsymbol{\gamma}_{\mathrm{CSPML}})} \\ \mathrm{S_i}(\boldsymbol{\gamma}_{\mathrm{I}}) \end{pmatrix}$

2. Apply the first order Taylor series expansion to $\mathrm{g}(\boldsymbol{\eta})$ at true values $(\boldsymbol{\gamma}_0, 0, \boldsymbol{\gamma}_0)^{\mathrm{T}}$:

$$\frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \begin{pmatrix} \mathrm{S_i}(\boldsymbol{\gamma}_0) \\ \mathrm{u_i}(\boldsymbol{\gamma}_0) \\ \mathrm{S_i}(\boldsymbol{\gamma}_0) \end{pmatrix} + \begin{pmatrix} -\mathrm{B} & \mathrm{C} & 0 \\ \mathrm{C} & \mathrm{L} & 0 \\ 0 & 0 & -\mathrm{B} \end{pmatrix} \begin{pmatrix} \boldsymbol{\gamma}_{\mathrm{CSPML}} - \boldsymbol{\gamma}_0 \\ \boldsymbol{\lambda} - 0 \\ \boldsymbol{\gamma}_{\mathrm{I}} - \boldsymbol{\gamma}_0 \end{pmatrix} + \mathrm{op}(1) = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$\Longleftrightarrow$

$$\sqrt{\mathrm{n}} \begin{pmatrix} \boldsymbol{\gamma}_{\mathrm{CSPML}} - \boldsymbol{\gamma}_0 \\ \boldsymbol{\lambda} - 0 \\ \boldsymbol{\gamma}_{\mathrm{I}} - \boldsymbol{\gamma}_0 \end{pmatrix} = - \begin{pmatrix} -\mathrm{B} & \mathrm{C} & 0 \\ \mathrm{C} & \mathrm{L} & 0 \\ 0 & 0 & -\mathrm{B} \end{pmatrix}^{-1} \frac{1}{\sqrt{\mathrm{n}}} \begin{pmatrix} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{S_i}(\boldsymbol{\gamma}_0) \\ \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{u_i}(\boldsymbol{\gamma}_0) \\ \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{S_i}(\boldsymbol{\gamma}_0) \end{pmatrix} + \mathrm{op}(1)$$

3. Asymptotic covariance under two conditions:

   - When the uncertainty of $\hat{\boldsymbol{\beta}}$ is ignorable (i.e. $\hat{\boldsymbol{\beta}}$ can be used to approximate the true $\boldsymbol{\beta}_0$)

$$\begin{aligned} \mathrm{Cov}(\hat{\boldsymbol{\eta}}) &= \frac{1}{\mathrm{n}} \begin{pmatrix} -\mathrm{B} & \mathrm{C} & 0 \\ \mathrm{C} & \mathrm{L} & 0 \\ 0 & 0 & -\mathrm{B} \end{pmatrix}^{-1} \frac{1}{\mathrm{n}} \mathrm{nE}\{ \begin{pmatrix} \mathrm{S_i}(\boldsymbol{\gamma}_0) \\ \mathrm{u_i}(\boldsymbol{\gamma}_0) \\ \mathrm{S_i}(\boldsymbol{\gamma}_0) \end{pmatrix} \begin{pmatrix} \mathrm{S_i}(\boldsymbol{\gamma}_0) \\ \mathrm{u_i}(\boldsymbol{\gamma}_0) \\ \mathrm{S_i}(\boldsymbol{\gamma}_0) \end{pmatrix}^{\mathrm{T}} \} \begin{pmatrix} -\mathrm{B} & \mathrm{C} & 0 \\ \mathrm{C} & \mathrm{L} & 0 \\ 0 & 0 & -\mathrm{B} \end{pmatrix}^{-1} \\ &= \frac{1}{\mathrm{n}} \begin{pmatrix} -\mathrm{B} & \mathrm{C} & 0 \\ \mathrm{C} & \mathrm{L} & 0 \\ 0 & 0 & -\mathrm{B} \end{pmatrix}^{-1} \begin{pmatrix} \mathrm{B} & 0 & \mathrm{B} \\ 0 & \mathrm{L} & 0 \\ \mathrm{B} & 0 & \mathrm{B} \end{pmatrix} \begin{pmatrix} -\mathrm{B} & \mathrm{C} & 0 \\ \mathrm{C} & \mathrm{L} & 0 \\ 0 & 0 & -\mathrm{B} \end{pmatrix}^{-1} \\ &= \frac{1}{\mathrm{n}} \begin{pmatrix} (\mathrm{B} + \mathrm{C}^{\mathrm{T}}\mathrm{L}^{-1}\mathrm{C})^{-1} & 0 & (\mathrm{B} + \mathrm{C}^{\mathrm{T}}\mathrm{L}^{-1}\mathrm{C})^{-1} \\ & \mathrm{B}(\mathrm{B} + \mathrm{C}^{\mathrm{T}}\mathrm{L}^{-1}\mathrm{C})^{-1} & -(\mathrm{C} + \mathrm{B}^{\mathrm{T}}\mathrm{C}^{-1}\mathrm{L})^{-1} \\ & & \mathrm{B}^{-1} \end{pmatrix} \end{aligned}$$

   - When the uncertainty of $\hat{\boldsymbol{\beta}}$ is not ignorable
     Assume that the external $\boldsymbol{\beta}$ estimator $\hat{\boldsymbol{\beta}}$ is derived from a finite sample of size N, whose variance is $\frac{\mathrm{V}_\beta}{\mathrm{N}}$. We incorporate the uncertainty from $\hat{\boldsymbol{\beta}}$ by applying the first order Taylor series expansion at the true $\boldsymbol{\beta}_0$ to the only $\boldsymbol{\beta}$-related parameter $u_i(\boldsymbol{\gamma}_0)$ in step 2. Let $\rho = \lim \frac{\mathrm{n}}{\mathrm{N}}$, and $\mathrm{Q} = \mathrm{E}\{\frac{\partial \mathrm{u_i}(\boldsymbol{\gamma}_0, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\}$

$$\frac{1}{\sqrt{\mathrm{n}}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{u_i}(\boldsymbol{\gamma}_0, \hat{\boldsymbol{\beta}}) = \frac{1}{\sqrt{\mathrm{n}}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{u_i}(\boldsymbol{\gamma}_0, \boldsymbol{\beta}_0) + \sqrt{\rho}\mathrm{Q}\sqrt{\mathrm{N}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

Thus,

$$
\begin{aligned}
\mathrm{Cov}(\hat{\boldsymbol{\eta}}) &= \frac{1}{\mathrm{n}}\begin{pmatrix}-\mathrm{B} & \mathrm{C} & 0\\ \mathrm{C} & \mathrm{L} & 0\\ 0 & 0 & -\mathrm{B}\end{pmatrix}^{-1}\begin{pmatrix}\mathrm{B} & 0 & \mathrm{B}\\ 0 & \mathrm{L}+\rho\mathrm{N}\mathrm{Q}^{\mathrm{T}}\mathrm{Var}(\hat{\boldsymbol{\beta}})\mathrm{Q} & 0\\ \mathrm{B} & 0 & \mathrm{B}\end{pmatrix}\begin{pmatrix}-\mathrm{B} & \mathrm{C} & 0\\ \mathrm{C} & \mathrm{L} & 0\\ 0 & 0 & -\mathrm{B}\end{pmatrix}^{-1}\\
&= \frac{1}{\mathrm{n}}\begin{pmatrix}-\mathrm{B} & \mathrm{C} & 0\\ \mathrm{C} & \mathrm{L} & 0\\ 0 & 0 & -\mathrm{B}\end{pmatrix}^{-1}\begin{pmatrix}\mathrm{B} & 0 & \mathrm{B}\\ 0 & \mathrm{L}+\rho\mathrm{Q}^{\mathrm{T}}\mathrm{V}_{\beta}\mathrm{Q} & 0\\ \mathrm{B} & 0 & \mathrm{B}\end{pmatrix}\begin{pmatrix}-\mathrm{B} & \mathrm{C} & 0\\ \mathrm{C} & \mathrm{L} & 0\\ 0 & 0 & -\mathrm{B}\end{pmatrix}^{-1}\\
&= \frac{1}{\mathrm{n}}\begin{pmatrix}{\scriptstyle \mathrm{Var}(\hat{\gamma}_{\mathrm{CSPML}})[1+\rho(\mathrm{C}^{\mathrm{T}}\mathrm{L}^{-1}\mathrm{Q}^{\mathrm{T}}\mathrm{V}_{\beta}\mathrm{Q}\mathrm{L}^{-1}\mathrm{C})\mathrm{Var}(\hat{\gamma}_{\mathrm{CSPML}})]} & \mathrm{term}_1 & {\scriptstyle \mathrm{Var}(\hat{\gamma}_{\mathrm{CSPML}})}\\ & \mathrm{term}_2 & -(\mathrm{C}+\mathrm{B}^{\mathrm{T}}\mathrm{C}^{-1}\mathrm{L})^{-1}\\ & & \mathrm{B}^{-1}\end{pmatrix}
\end{aligned}
$$

$$
\text{where}\begin{cases}\mathrm{Var}(\hat{\boldsymbol{\gamma}}_{\mathrm{CSPML}}) = (\mathrm{B}+\mathrm{C}^{\mathrm{T}}\mathrm{L}^{-1}\mathrm{C})^{-1},\\ \mathrm{term}_1 = \rho\mathrm{Var}(\hat{\boldsymbol{\gamma}}_{\mathrm{CSPML}})(\mathrm{B}\mathrm{L}^{-1}\mathrm{Q}^{\mathrm{T}}\mathrm{V}_{\beta}\mathrm{Q}\mathrm{L}^{-1}\mathrm{C})\mathrm{Var}(\hat{\boldsymbol{\gamma}}_{\mathrm{CSPML}}),\\ \mathrm{term}_2 = \mathrm{B}\mathrm{Var}(\hat{\boldsymbol{\gamma}}_{\mathrm{CSPML}}) + \mathrm{B}\mathrm{Var}(\hat{\boldsymbol{\gamma}}_{\mathrm{CSPML}})(\rho\mathrm{Q}^{\mathrm{T}}\mathrm{V}_{\beta}\mathrm{Q})\mathrm{Var}(\hat{\boldsymbol{\gamma}}_{\mathrm{CSPML}})\mathrm{B}.\end{cases}
$$

B.3 Re-parameterize $\hat{\boldsymbol{\gamma}}_{\mathrm{EB}}$

Based on matrix 3.2 in *Proposition 1* of Section 3.2.4, we can show that

$$
\begin{cases}\hat{\boldsymbol{\gamma}}_{\mathrm{I}} \mid \boldsymbol{\gamma} \sim \mathrm{N}(\boldsymbol{\gamma}, \hat{\Sigma})\\ \hat{\boldsymbol{\gamma}}_{\mathrm{CSPML}} \mid \boldsymbol{\gamma} \sim \mathrm{N}(\boldsymbol{\gamma}, \hat{\mathrm{V}}_{\mathrm{CSPML}})\end{cases} \Longrightarrow \begin{pmatrix}\mathrm{Z}\\ \hat{\boldsymbol{\gamma}}_{\mathrm{CSPML}}\\ \hat{\boldsymbol{\gamma}}_{\mathrm{I}}\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}0\\ \boldsymbol{\gamma}\\ \boldsymbol{\gamma}\end{pmatrix}, \begin{pmatrix}\hat{\mathrm{V}}_{\mathrm{Z}} \equiv \hat{\mathrm{V}}_{\mathrm{I}} - \hat{\mathrm{V}}_{\mathrm{CSPML}} & 0 & \hat{\mathrm{V}}_{\mathrm{Z}}\\ 0 & \hat{\mathrm{V}}_{\mathrm{CSPML}} & \hat{\mathrm{V}}_{\mathrm{CSPML}}\\ \hat{\mathrm{V}}_{\mathrm{Z}} & \hat{\mathrm{V}}_{\mathrm{CSPML}} & \hat{\mathrm{V}}_{\mathrm{I}}\end{pmatrix}\right).
$$

$\mathrm{Cov}(\mathrm{Z}, \hat{\boldsymbol{\gamma}}_{\mathrm{CSPML}}) = 0$ can also be derived from the conclusion that $\mathrm{Cov}(\hat{\boldsymbol{\gamma}}_{\mathrm{I}}, \hat{\boldsymbol{\gamma}}_{\mathrm{CSPML}}) = \mathrm{Var}(\hat{\boldsymbol{\gamma}}_{\mathrm{CSPML}})$ in *Proposition 1*. Thus, we can use Z, the scalar $\mathrm{Z}^{\mathrm{T}}\mathrm{V}_{\mathrm{I}}^{-1}\mathrm{Z}$ and Woodbury formula [Golub and Loan, 1996] to re-parameterize $\hat{\boldsymbol{\gamma}}_{\mathrm{EB}}$:

$$
\begin{aligned}
\hat{\boldsymbol{\gamma}}_{\mathrm{EB}} &= \hat{\mathrm{A}}(\hat{\Sigma}+\hat{\mathrm{A}})^{-1}\hat{\boldsymbol{\gamma}}_{\mathrm{I}} + \hat{\Sigma}(\hat{\Sigma}+\hat{\mathrm{A}})^{-1}\hat{\boldsymbol{\gamma}}_{\mathrm{CSPML}}\\
&= \mathrm{Z}\mathrm{Z}^{\mathrm{T}}(\hat{\mathrm{V}}_{\mathrm{I}}+\mathrm{Z}\mathrm{Z}^{\mathrm{T}})^{-1}(\mathrm{Z}+\hat{\boldsymbol{\gamma}}_{\mathrm{CSPML}}) + \hat{\mathrm{V}}_{\mathrm{I}}(\hat{\mathrm{V}}_{\mathrm{I}}+\mathrm{Z}\mathrm{Z}^{\mathrm{T}})^{-1}\hat{\boldsymbol{\gamma}}_{\mathrm{CSPML}}\\
&= \hat{\boldsymbol{\gamma}}_{\mathrm{CSPML}} + \mathrm{Z}\mathrm{Z}^{\mathrm{T}}(\hat{\mathrm{V}}_{\mathrm{I}}+\mathrm{Z}\mathrm{Z}^{\mathrm{T}})^{-1}\mathrm{Z}\\
&= \hat{\boldsymbol{\gamma}}_{\mathrm{CSPML}} + \mathrm{Z}(1-\frac{1}{1+\mathrm{Z}^{\mathrm{T}}\mathrm{V}_{\mathrm{I}}^{-1}\mathrm{Z}}).
\end{aligned}
$$

Or equivalently, $\hat{\boldsymbol{\gamma}}_{\mathrm{EB}} = \hat{\boldsymbol{\gamma}}_{\mathrm{I}} - \hat{\mathrm{V}}_{\mathrm{I}}(\hat{\mathrm{V}}_{\mathrm{I}}+\mathrm{Z}\mathrm{Z}^{\mathrm{T}})^{-1}\mathrm{Z} = \hat{\boldsymbol{\gamma}}_{\mathrm{I}} - \mathrm{Z}\frac{1}{1+\mathrm{Z}^{\mathrm{T}}\mathrm{V}_{\mathrm{I}}^{-1}\mathrm{Z}}$.

B.4 Additional Simulation Results

Figure B.2 summarizes the additional simulation settings we assessed regarding simulation III in Section 3.3. The goal is to evaluate the performance of proposed approach and the comparisons when the joint distribution of (X, B) is misspecified in the external model 1. Scenario (1) is the same as simulation III in the main manuscript; scenario (2) assesses different (X, B) in external model 1 through the different conditional distribution $B|X$; and scenario (3) varies the joint distribution of (X, B) through the marginal distribution of X.

	Internal Data	External Model		
		1	2	3
Model	Y|X1, X2, X3, X4, B	Y|X1, X2	Y|X1, X2	Y|X1, X2, X3, X4
The joint distribution of (X, B)	(X, B) ~N ((0,0), 1), correlation 0.3 X ~ N (0, 1), correlation 0.3; B|X ~ (-1+0.5X, 1)	(1) (X, B) is different from the marginal distribution of B: (X, B) ~ N ((0, 1.5), 1), correlation 0.3	Same as the internal data	Same as the internal data
		(2) (X, B) is different from the conditional distribution B|X: X ~ N (0, 1), correlation 0.3; B|X ~ (-1+0.25X, 1)		
		(3) (X, B) is different from the marginal distribution X: X ~ N (0.25, 1), correlation 0.3; B|X ~ (-1+0.5X, 1)		
Sample size	n=200	m_1=30,000	m_2=30,000	m_3=30,000

Figure B.2: Simulation settings of additional scenarios for simulation III in Chapter 3: the joint distribution of (X, B) is misspecified in the external model 1.

For simulation IV in Section 3.3, we have additionally assessed different forms of misspecified outcome model in the external population listed in Figure B.3, including different intercept in scenario (1), different X coefficients in scenario (2), and different intercept and X coefficients in scenario (3). Scenario (3) is the same as simulation IV in the main manuscript.

	Internal Data	External Model		
		1	2	3
Model	Y|X1, X2, X3, X4, B	Y|X1, X2	Y|X1, X3	Y|X1, X2, X3, X4
Outcome model	logit Pr(Y=1|X, B)= -1-0.5X+0.5B	Same as the internal data	Same as the internal data	(1) intercept different: logit Pr(Y=1|X, B)= 1-0.5X+0.5B (2) X coefficients different: logit Pr(Y=1|X, B)= -1+0.5X+0.5B (3) intercept & X coefficients different: logit Pr(Y=1|X, B)= 1+0.5X+0.5B
Sample size	n=200	m1=30,000	m2=30,000	m3=30,000

Note: (X, B) followed Gaussian distribution with mean zero, standard deviation 1, and correlation 0.3.

Figure B.3: Simulation settings of additional scenarios for simulation IV in Chapter 3: the outcome model is misspecified in the external model 3.

The results in Figure B.4 corresponding to Figure B.2 indicate that CSPML estimators would have substantial bias when the difference comes from the conditional distribution $\mathrm{B}|\mathrm{X}$ or marginal distribution B, but not troublesome when the marginal distribution X is different. In the cases where CSPML estimators are biased, the EB estimator will protect against the bias by downweighting the CSPML estimator and sacrificing some precision. In summary, these simulation results provided the evidence that the proposed approach is robust to the heterogeneous covariate distribution (X, B) of the external population.

The result in Figure B.5 corresponding to Figure B.3 indicates that when the misspecification gets worse (from misspecified intercept only to both misspecified intercept and X coefficients), the EB estimator will trade more efficiency gain to correct the bias from the CSPML estimator and thus we obtain less efficiency by incorporating this external model (OCWE assigned smaller weights on the corresponding estimator EB_3). More specifically, when we had mild misspecification in the external model 3 (only the intercept, i.e., the prevalence in the external population is different from the internal population), the external model 3 can still contribute 40% in the final OCWE. However, when the misspecification gets worse (both intercept and X coefficients are misspecified), the OCWE basically rejects the wrong information from the external model 3 by incorporating only 4% of it.

B.5 External Calculators in Prostate Cancer Data Example

In the real data analysis in Section 3.4, we incorporate two external risk calculators of high-grade prostate cancer, the Prostate Cancer Prevention Trial (PCPThg) [Thompson et al., 2006] and the European Randomized Study of Screening for Prostate Cancer (ERSPC) risk calculator 3 [Roobol

		Direct Regression	Internal Data + External 1		Internal Data + External 2		Internal Data + External 3		Composite of EB Estimators		
			CSPML 1	EB 1	CSPML 2	EB 2	CSPML 3	EB 3	IVW	OCWE	SC-Learner
		Bias (SD) [ESE] 95% Coverage Rate									
(1)	Weights	/	/	/	/	/	/	/	[.33, .33, .33]	[.12, .38, .49]	/
	γ_0	-.042 (.198) [.193] 95%	.665 (.080) [.102] 0%	-.001 (.201) [.162] 90%	-.039 (.081) [.085] 97%	-.040 (.163) [.161] 95%	-.033 (.051) [.058] 0%	-.039 (.169) [.169] 93%	-.027 (.177) [.163] 93%	-.031 (.175) [.163] 93%	-.027 (.177) [.163] 93%
	γ_1	-.020 (.229) [.208] 94%	.02 (.094) [.112] 98%	-.018 (.219) [.175] 89%	.012 (.095) [.097] 97%	-.011 (.190) [.174] 93%	.0220(.066) [.074] 100%	-.013 (.198) [.182] 90%	-.014 (.202) [.176] 92%	-.012 (.197) [.177] 93%	-.014 (.202) [.176] 92%
	γ_2	-.018 (.220) [.208] 94%	.004 (.090) [.112] 98%	-.016 (.210) [.175] 90%	-.015 (.091) [.096] 97%	-.015 (.180) [.173] 94%	.005 (.062) [.073] 100%	-.015 (.188) [.182] 91%	-.016 (.192) [.176] 93%	-.015 (.187) [.177] 94%	-.016 (.199) [.177] 93%
	γ_3	-.019 (.208) [.207] 96%					-.014 (.062) [.072] 100%	-.017 (.178) [.194] 93%	-.018 (.198) [.195] 96%	-.017 (.193) [.192] 96%	-.018 (.192) [.192] 96%
	γ_4	-.027 (.217) [.208] 94%					-.021 (.067) [.073] 97%	-.025 (.188) [.195] 94%	-.026 (.207) [.196] 94%	-.025 (.202) [.193] 94%	-.025 (.188) [.195] 96%
	γ_B	.017 (.219) [.207] 98%									
(2)	Weights	/	/	/	/	/	/	/	[.33, .33, .33]	[.26, .26, .47]	/
	γ_0	-.025 (.369) [.251] 95%	.055 (.167) [.164] 94%	-.009 (.243) [.222] 94%	-.025 (.167) [.163] 95%	-.023 (.235) [.222] 96%	-.013 (.160) [.154] 95%	-.021 (.244) [.229] 93%	-.017 (.240) [.224] 95%	-.017 (.240) [.225] 95%	-.017 (.240) [.224] 95%
	γ_1	-.049 (.244) [.39] 94%	-.169 (.140) [.143] 83%	-.007 (.221) [.208] 93%	-.010 (.139) [.142] 95%	-.038 (.212) [.208] 94%	.007 (.122) [.127] 95%	-.038 (.217) [.216] 94%	-.049 (.210) [.210] 94%	-.047 (.211) [.211] 94%	-.049 (.210) [.210] 94%
	γ_2	-.041 (.255) [.240] 94%	-.174 (.147) [.144] 82%	-.065 (.231) [.209] 93%	-.041 (.146) [.142] 95%	-.041 (.223) [.209] 93%	-.029 (.127) [.127] 96%	-.039 (.228) [.216] 94%	-.048 (.210) [.210] 94%	-.047 (.211) [.211] 94%	-.052 (.211) [.211] 94%
	γ_3	-.023 (.255) [.240] 93%					-.017 (.126) [.126] 96%	-.022 (.227) [.228] 95%	-.022 (.210) [.229] 93%	-.022 (.211) [.226] 93%	-.022 (.211) [.226] 93%
	γ_4	-.014 (.232) [.238] 97%					-.055 (.124) [.126] 96%	-.021 (.206) [.227] 97%	-.016 (.223) [.228] 96%	-.017 (.219) [.225] 96%	-.021 (.206) [.227] 97%
	γ_B	.037 (.211) [.202] 98%									
(3)	Weights	/	/	/	/	/	/	/	[.33, .33, .33]	[.26, .27, .47]	/
	γ_0	-.025 (.369) [.251] 95%	-.061 (.167) [.163] 94%	-.030 (.235) [.222] 95%	-.025 (.167) [.163] 95%	-.023 (.235) [.222] 96%	-.013 (.160) [.154] 95%	-.021 (.244) [.229] 93%	-.025 (.238) [.224] 95%	-.024 (.239) [.225] 95%	-.025 (.238) [.224] 95%
	γ_1	-.049 (.244) [.39] 94%	-.008 (.139) [.142] 95%	-.037 (.212) [.208] 94%	-.010 (.139) [.142] 95%	-.038 (.212) [.208] 94%	.007 (.122) [.127] 95%	-.038 (.217) [.216] 94%	-.038 (.213) [.210] 94%	-.038 (.214) [.211] 94%	-.038 (.213) [.210] 94%
	γ_2	-.041 (.255) [.240] 94%	-.030 (.146) [.142] 95%	-.038 (.224) [.240] 93%	-.041 (.146) [.142] 95%	-.041 (.223) [.209] 93%	-.029 (.127) [.127] 96%	-.039 (.228) [.216] 94%	-.039 (.224) [.210] 94%	-.039 (.225) [.211] 94%	-.038 (.225) [.211] 94%
	γ_3	-.023 (.255) [.240] 93%					-.017 (.126) [.126] 96%	-.022 (.227) [.228] 95%	-.022 (.245) [.229] 93%	-.022 (.241) [.226] 93%	-.022 (.240) [.226] 93%
	γ_4	-.014 (.232) [.238] 97%					-.055 (.124) [.126] 96%	-.021 (.206) [.227] 97%	-.016 (.223) [.228] 96%	-.016 (.220) [.225] 95%	-.021 (.206) [.227] 97%
	γ_B	.037 (.211) [.202] 98%									

Abbreviations: CSPML, constrained maximal likelihood estimator; EB, empirical Bayes estimator; SD, Monte Carlo standard deviation from 500 simulations; ESE, estimated average standard error from 500 simulations; IVW, Inverse variance-weighted estimator; OCWE, Optimal covariance-weighted estimator; SC-Learner, Selective coefficient learner. **Note**: Internal dataset size n=200; green represents good performance with small bias, yellow represents overestimated ESE (in square brackets) compared with the Monte Carlo SD (in round brackets), and red represents the poor performance of bias/95% coverage rate.

Figure B.4: Simulation results of additional scenarios of simulation III in Chapter 3.

	Direct Regression	Internal Data + External 1		Internal Data + External 2			Internal Data + External 3		Composite of EB Estimators		
		CSPML 1	EB 1	CSPML 2	EB 2		CSPML 3	EB 3	IVW	OCWE	SC-Learner
	Bias (SD) [ESE] 95% Coverage Rate										
						weight	/	/	[.33, .33, .33]	[.41, .17, .41]	/
Y_0	-.041 (.198) [.193] 95%	-.008 (.081) [.085] 96%	-.032 (.163) [.161] 94%	-.007 (.089) [.084] 95%	-.032 (.164) [.161] 95%		1.89 (.054) [.142] 0%	-.026 (.198) [.174] 93%	-.034 (.174) [.164] 93%	-.029 (.177) [.165] 93%	-.035 (.173) [.163] 93%
Y_1	-.020 (.230) [.208] 94%	-.005 (.095) [.097] 96%	-.014 (.190) [.174] 93%	-.002 (.089) [.097] 97%	-.014 (.189) [.174] 93%		.006 (.065) [.149] 100%	-.019 (.227) [.186] 90%	-.020 (.201) [.177] 90%	-.018 (.203) [.177] 90%	-.020 (.200) [.176] 90%
Y_2	-.018 (.220) [.208] 95%	-.008 (.091) [.096] 97%	-.014 (.180) [.175] 94%			(1)	.010 (.064) [.149] 100%	-.017 (.218) [.187] 91%	-.017 (.205) [.187] 91%	-.014 (.201) [.182] 91%	-.017 (.197) [.175] 91%
Y_3	-.019 (.208) [.207] 96%			-.003 (.089) [.095] 98%	-.013 (.171) [.180] 96%		-.012 (.064) [.147] 100%	-.019 (.207) [.185] 93%	-.019 (.195) [.186] 93%	-.016 (.200) [.190] 93%	-.018 (.188) [.178] 93%
Y_4	-.026 (.216) [.208] 94%						-.012 (.066) [.148] 97%	-.026 (.214) [.189] 94%	-.026 (.216) [.200] 94%	-.026 (.215) [.198] 94%	-.026 (.214) [.189] 94%
Y_B	.019 (.218) [.207] 98%					Y_B					
						weight	/	/	[.34, .33, .32]	[.54, .33, .13]	/
						Y_0	-.03 (.089) [.153] 99%	-.04 (.196) [.175] 93%	-.038 (.173) [.164] 94%	-.035 (.165) [.161] 94%	-.038 (.172) [.163] 94%
						Y_1	.858 (.137) [.164] 0%	-.013 (.227) [.189] 90%	-.017 (.200) [.177] 92%	-.016 (.193) [.174] 93%	-.017 (.199) [.177] 92%
						Y_2	.864 (.136) [.165] 0%	-.009 (.216) [.189] 92%	-.013 (.203) [.188] 93%	-.012 (.195) [.186] 94%	-.012 (.195) [.176] 93%
						(2) Y_3	.855 (.138) [.163] 1%	-.013 (.208) [.188] 94%	-.016 (.195) [.187] 95%	-.013 (.196) [.190] 95%	-.016 (.188) [.180] 95%
						Y_4	.875 (.139) [.164] 1%	-.021 (.215) [.191] 93%	-.025 (.216) [.200] 94%	-.026 (.216) [.205] 94%	-.021 (.215) [.191] 93%
						Y_B					
						weight	/	/	[.34, .33, .32]	[.57, .38, .04]	/
						Y_0	1.69 (.189) [.162] 0%	-.031 (.169) [.178] 96%	-.032 (.173) [.165] 93%	-.030 (.164) [.160] 93%	-.032 (.172) [.164] 93%
						Y_1	.747 (.178) [.177] 3%	-.016 (.197) [.192] 95%	-.015 (.201) [.178] 92%	-.013 (.191) [.173] 93%	-.015 (.200) [.178] 92%
						Y_2	.756 (.174) [.177] 3%	-.014 (.188) [.192] 94%	-.015 (.206) [.189] 93%	-.013 (.196) [.187] 94%	-.014 (.197) [.177] 93%
						(3) Y_3	.762 (.167) [.175] 1%	-.014 (.178) [.191] 95%	-.015 (.194) [.187] 95%	-.013 (.193) [.189] 94%	-.014 (.187) [.181] 95%
						Y_4	.746 (.179) [.176] 4%	-.022 (.188) [.193] 93%	-.025 (.216) [.202] 94%	-.026 (.217) [.207] 95%	-.022 (.216) [.193] 94%
						Y_B					

Abbreviations: CSPML, constrained maximal likelihood estimator; EB, empirical Bayes estimator; SD, Monte Carlo standard deviation from 500 simulations; ESE, estimated average standard error from 500 simulations; IVW, Inverse variance-weighted estimator; OCWE, Optimal covariance-weighted estimator; SC-Learner, Selective coefficient learner. **Note**: Internal dataset size n=200; green represents good performance with small bias, yellow represents overestimated ESE (in square brackets) compared with the Monte Carlo SD (in round brackets), and red represents the poor performance of bias/95% coverage rate.

Figure B.5: Simulation results of additional scenarios of simulation IV in Chapter 3.

et al., 2012].

The proposed framework requires the consistent type of the common parameters shared across different external models (e.g. if one external model used mean-centered continuous age as a predictor, then all other external models including the target outcome model need to use the same age variable). Since PCPThg used log-based PSA and ERSPC used mean-centered log2-based PSA, we need to reconcile the different transformation by adjusting the reported intercept. In PCPThg, the authors only reported median PSA [Thompson et al., 2006] while ERSPC reported both median and mean PSA in their Table 1 [Roobol et al., 2012]. Therefore, we decided to use median-centered log2-based PSA, mean-centered continuous age, and mean-centered biopsy variable throughout.

We adjusted the originally reported estimated coefficients and intercept as follows:

- Original PCPThg: $\mathrm{logit}(\mathrm{p_i}) = -6.25 + 1.29\mathrm{log}(\mathrm{PSA_i}) + \mathrm{DRE_i} + 0.03\mathrm{Age_i} - 0.36\mathrm{Biopsy_i} + 0.96\mathrm{Race_i}$;
- $\hat{\beta}_{\log_2(\mathrm{PSA})} = \frac{\hat{\beta}_{\log(\mathrm{PSA})}}{\log_2(\mathrm{e})} = \frac{1.29}{\log_2(\mathrm{e})} = 0.8941599$;
- Adjusted $\hat{\beta}_0 = \hat{\beta}_0 + \hat{\beta}_{\log_2(\mathrm{PSA})} \times [\text{median } \log_2(\mathrm{PSA})] + \hat{\beta}_{\mathrm{Age}}\overline{\mathrm{Age}} + \hat{\beta}_{\mathrm{Biopsy}}\overline{\mathrm{Biopsy}}$
  $= -6.25 + 0.8941599 \times \log_2(1.5) + 0.03 \times 69.66 - 0.36 \times (753/5519)$
  $= -3.686268$

where median PSA can be found in the first paragraph of Results, and mean age (i.e. $\overline{\mathrm{Age}}$) and mean biopsy (i.e. $\overline{\mathrm{Biopsy}}$) can be estimated from Table 2 in Thompson et al. [2006].

Similarly, since the $\log_2(\mathrm{PSA})$ in the original ERSPC risk calculator was mean-centered, we need to transform it to median-centered by adjusting the intercept as follows:

- Original ERSPC: $\mathrm{logit}(\mathrm{p_i}) = -3.51 + 1.18\log_2(\mathrm{PSA_i}) + 1.81\mathrm{DRE_i} - 1.51\log_2(\mathrm{TRUS} - \mathrm{PV_i})$
- Adjusted $\hat{\beta}_0 = \hat{\beta}_0 + \hat{\beta}_{\log_2(\mathrm{PSA})} \times [\text{median } \log_2(\mathrm{PSA}) - \overline{\log(\mathrm{PSA})}]$
  $\approx \hat{\beta}_0 + \hat{\beta}_{\log_2(\mathrm{PSA})} \times [\log_2(\text{median PSA}) - \log(\overline{\mathrm{PSA}})]$
  $= -3.51 + 1.18 \times [\log_2(4.3) - \log(6.1)]$
  $= -3.16$

where mean and median PSA can be found in Table 1 from Roobol et al. [2012].

For other estimated coefficients that are not mentioned here, we used the originally reported values in the analysis.

APPENDIX C

Appendix of Chapter 4

C.1 Deriving the Initial Estimates for the External Populations

In this section, we will show how to obtain the initial parameter estimates of external population k. Let $(\hat{\gamma}_0^{\mathrm{S_0}}, \hat{\boldsymbol{\gamma}}_{\mathrm{X}}^{\mathrm{S_0T}}, \hat{\boldsymbol{\gamma}}_{\mathrm{B}}^{\mathrm{S_0T}})^{\mathrm{T}}$ be the direct regression estimates of $\mathrm{Y}|\mathbf{X}, \mathbf{B}, \mathrm{S} = 0$ using internal data only. For external population k, we know the parameter estimates $\hat{\boldsymbol{\beta}}_{\mathrm{k}} = (\hat{\beta}_0, \hat{\boldsymbol{\beta}}_{\mathrm{X}}^{\mathrm{T}})^{\mathrm{T}}$ from the fitted model $\mathrm{Y}|\mathbf{X}_{\mathrm{k}}; \boldsymbol{\beta}_{\mathrm{k}}$. We assume that all predictors, $\mathbf{X}$ and $\mathbf{B}$, are centered, and the true target model parameter for the external population k is $(\gamma_0^{\mathrm{S_k}}, \boldsymbol{\gamma}_{\mathrm{X}}^{\mathrm{S_kT}}, \boldsymbol{\gamma}_{\mathrm{B}}^{\mathrm{S_0T}})^{\mathrm{T}}$, assuming the coefficient of the unobserved variable $\mathbf{B}$ is the same as the internal population, i.e. $\boldsymbol{\gamma}_{\mathrm{B}}^{\mathrm{S_k}} = \boldsymbol{\gamma}_{\mathrm{B}}^{\mathrm{S_0}}$.

The goal of estimating $\gamma_0^{\mathrm{S_k}}$ and $\boldsymbol{\gamma}_{\mathrm{X}}^{\mathrm{S_k}}$ from model $\mathrm{Y}|\mathbf{X}, \mathbf{B}, \mathrm{S} = \mathrm{k}; \boldsymbol{\gamma}^{\mathrm{S_k}}$ is equivalent to correcting the bias of $\hat{\boldsymbol{\beta}}_{\mathrm{k}}$ in the reduced model $\mathrm{Y}|\mathbf{X}_{\mathrm{k}}; \boldsymbol{\beta}_{\mathrm{k}}$ considering covariates $\mathbf{X}_{(-\mathrm{k})}$ and $\mathbf{B}$ as omitted. To simplify notation, we assume $\mathbf{B}$ is the only omitted covariate in the derivation below. Neuhaus and Jewell [1993] provided a Taylor-series-expansion approximation to show that the ratio of coefficients remains constant in both the reduced and the full model when the omitted $\mathbf{B}$ is independent of the observed $\mathbf{X}$, i.e. $\frac{\gamma_{\mathrm{X_1}}}{\gamma_{\mathrm{X_2}}} \approx \frac{\beta_{\mathrm{X_1}}}{\beta_{\mathrm{X_2}}}$, indicating that the relative effect size among regression coefficients remains consistent across models. In their Table 3 and equation 9, Neuhaus and Jewell [1993] provided the algebraic relationship between $\boldsymbol{\gamma}_{\mathrm{X}}^{\mathrm{S_k}}$ and $\boldsymbol{\beta}_{\mathrm{X}}$ for exponential family when the omitted $\mathbf{B}$ and the observed $\mathbf{X}$ are correlated. In the subsequent paragraphs, we will explain in detail how to estimate $\gamma_0^{\mathrm{S_k}}$ and $\boldsymbol{\gamma}_{\mathrm{X}}^{\mathrm{S_k}}$ in linear regression (continuous Y) and logistic regression (binary Y), respectively.

1. Linear Regression: Suppose $\mathrm{E}(\mathbf{B}|\mathbf{X}; \boldsymbol{\theta}) = \boldsymbol{\theta}\mathrm{X}$. We start by replacing $\mathbf{B}$ with the conditional expected value $\mathrm{E}(\mathbf{B}|\mathbf{X}; \boldsymbol{\theta})$ in the mean profile of the target model:

$$\mathrm{E}(\mathrm{Y}|\mathbf{X}, \mathbf{B}; \boldsymbol{\gamma}) = \gamma_0^{\mathrm{S_k}} + \boldsymbol{\gamma}_{\mathrm{X}}^{\mathrm{S_k}}\mathbf{X} + \boldsymbol{\gamma}_{\mathrm{B}}^{\mathrm{S_0}}\mathbf{B} = \gamma_0^{\mathrm{S_k}} + \boldsymbol{\gamma}_{\mathrm{X}}^{\mathrm{S_k}}\mathbf{X} + \boldsymbol{\gamma}_{\mathrm{B}}^{\mathrm{S_0}}\boldsymbol{\theta}\mathbf{X} = \mathrm{E}(\mathrm{Y}|\mathbf{X}; \boldsymbol{\gamma}, \boldsymbol{\theta})$$

Since $\hat{\mathrm{E}}(\mathrm{Y}|\mathbf{X}; \boldsymbol{\beta}) = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}_{\mathrm{X}}\mathbf{X}$ is available through the externally fitted model, we can obtain the estimation of $\gamma_0^{\mathrm{S_k}}$ and $\boldsymbol{\gamma}_{\mathrm{X}}^{\mathrm{S_k}}$ by matching the intercept and $\mathbf{X}$ coefficient between $\hat{\mathrm{E}}(\mathrm{Y}|\mathbf{X}; \boldsymbol{\gamma}, \boldsymbol{\theta})$ and $\hat{\mathrm{E}}(\mathrm{Y}|\mathbf{X}; \boldsymbol{\beta})$, respectively: $\hat{\gamma}_0^{\mathrm{S_k}} = \hat{\beta}_0$ and $\hat{\boldsymbol{\gamma}}_{\mathrm{X}}^{\mathrm{S_k}} = \hat{\boldsymbol{\beta}}_{\mathrm{X}} - \boldsymbol{\theta}^{\mathrm{T}}\hat{\boldsymbol{\gamma}}_{\mathrm{B}}^{\mathrm{S_0}}$. In a special case where the internal and

the external population only differ in intercept and S is independent of $\mathbf{X}$ and $\mathbf{B}$, we can directly obtain the initial estimates $\hat{\boldsymbol{\gamma}}_{\mathrm{k}} = (\hat{\beta}_0, \hat{\boldsymbol{\gamma}}_{\mathrm{X}}^{\mathrm{S_k T}}, \hat{\boldsymbol{\gamma}}_{\mathrm{B}}^{\mathrm{S_0 T}})^{\mathrm{T}}$.

2. Logistic Regression: In logistic regression where g() is the logit link function, we connect the intercepts β_0 and $\gamma_0^{\mathrm{S_k}}$ through the equation $\mathrm{logit}^{-1}(\beta_0) = \mathrm{E}_{\mathrm{B|X}}(\mu_0^{\mathrm{S_k}})$, where $\mu_0^{\mathrm{S_k}} = \mathrm{g}^{-1}(\mathrm{Y}|\mathbf{X}, \mathbf{B}; \boldsymbol{\gamma}_{\mathrm{X}}^{\mathrm{S_k}} = 0) = \mathrm{logit}^{-1}(\gamma_0^{\mathrm{S_k}} + \mathbf{B}^{\mathrm{T}}\boldsymbol{\gamma}_{\mathrm{B}}^{\mathrm{S_0}})$. For the right hand side, we expand $\mathbf{B}$, a vector of length Q, at $\mathrm{E}(\mathbf{B}|\mathbf{X})$ using the third-order Taylor series expansion as follows:

$$
\begin{aligned}
\mathrm{E}_{\mathrm{B|X}}(\mu_0^{\mathrm{S_k}}) &= \mathrm{E}_{\mathrm{B|X}}[\mathrm{logit}^{-1}(\gamma_0^{\mathrm{S_k}} + \mathbf{B}^{\mathrm{T}}\boldsymbol{\gamma}_{\mathrm{B}}^{\mathrm{S_0}})] \\
&\approx \mathrm{logit}^{-1}(\mathrm{w})\Big\{1 + \frac{1}{2}\frac{1-\mathrm{e}^{\mathrm{w}}}{(1+\mathrm{e}^{\mathrm{w}})^2}\sum_{\mathrm{i}=1}^{\mathrm{Q}}\sum_{\mathrm{j}=1}^{\mathrm{Q}}\gamma_{\mathrm{B_i}}^{\mathrm{S_0}}\gamma_{\mathrm{B_j}}^{\mathrm{S_0}}\mathrm{E}_{\mathrm{B|X}}\Big[\Big(\mathrm{B_i} - \mathrm{E}(\mathrm{B_i}|\mathbf{X})\Big)\Big(\mathrm{B_j} - \mathrm{E}(\mathrm{B_j}|\mathbf{X})\Big)\Big]\Big\} \\
&= \mathrm{logit}^{-1}(\mathrm{w})\Big[1 + \frac{1}{2}\frac{1-\mathrm{e}^{\mathrm{w}}}{(1+\mathrm{e}^{\mathrm{w}})^2}\mathrm{Var}\Big(\sum_{\mathrm{i}=1}^{\mathrm{Q}}\gamma_{\mathrm{B_i}}^{\mathrm{S_0}}\mathrm{B_i}|\mathbf{X}\Big)\Big] \qquad \text{(C.1)}
\end{aligned}
$$

where $\mathrm{w} = \gamma_0^{\mathrm{S_k}} + \mathrm{E}(\mathbf{B}^{\mathrm{T}}|\mathbf{X})\boldsymbol{\gamma}_{\mathrm{B}}^{\mathrm{S_0}}$. Given $\hat{\beta}_0, \hat{\boldsymbol{\gamma}}_{\mathrm{B}}^{\mathrm{S_0}}, \hat{\mathrm{E}}(\mathbf{B}|\mathbf{X})$ and $\hat{\mathrm{V}}\mathrm{ar}(\mathbf{B}|\mathbf{X})$, we can easily obtain $\hat{\gamma}_0^{\mathrm{S_k}}$ by solving the equation $\mathrm{E}_{\mathrm{B|X}}(\mu_0^{\mathrm{S_k}}) - \mathrm{logit}^{-1}(\hat{\beta}_0) = 0$.

After obtaining $\boldsymbol{\gamma}_0^{\mathrm{S_k}}$, we then estimate $\boldsymbol{\gamma}_{\mathrm{X}}^{\mathrm{S_k}} = (\gamma_{\mathrm{X_1}}^{\mathrm{S_k}}, ..., \gamma_{\mathrm{X_{P_k}}}^{\mathrm{S_k}})^{\mathrm{T}}$ according to the following equation provided in Neuhaus and Jewell [1993]:

$$
\boldsymbol{\beta}_{\mathrm{X_p}} = \Big\{\boldsymbol{\gamma}_{\mathrm{X_p}}^{\mathrm{S_k}} + \Big[\mathrm{E}(\mathbf{B}^{\mathrm{T}}|\mathbf{X} + 1_{\mathrm{p}}) - \mathrm{E}(\mathbf{B}^{\mathrm{T}}|\mathbf{X})\Big]\boldsymbol{\gamma}_{\mathrm{B}}^{\mathrm{S_0}}\Big\}\Big\{1 - \frac{\mathrm{Var}_{\mathrm{B|X}}(\mu_0^{\mathrm{S_k}})}{1 - \mathrm{E}_{\mathrm{B|X}}(\mu_0^{\mathrm{S_k}})[1 - \mathrm{E}_{\mathrm{B|X}}(\mu_0^{\mathrm{S_k}})]}\Big\}
$$

where 1_{p} is a zero vector with the p^{th} term equals to 1 and $\mathrm{p} \in \{1, ..., \mathrm{P_k}\}$. Similar to equation C.1, we can also obtain the Taylor-series-expansion estimation for $\mathrm{E}_{\mathrm{B|X}}[(\mu_0^{\mathrm{S_k}})^2] = \mathrm{E}_{\mathrm{B|X}}[\mathrm{logit}^{-2}(\gamma_0^{\mathrm{S_k}} + \mathbf{B}^{\mathrm{T}}\boldsymbol{\gamma}_{\mathrm{B}}^{\mathrm{S_0}})] \approx \frac{\mathrm{e}^{2\mathrm{w}}}{(1+\mathrm{e}^{\mathrm{w}})2}\Big[1 + \frac{1}{2}\frac{2-\mathrm{e}^{\mathrm{w}}}{(1+\mathrm{e}^2)^2}\sum_{\mathrm{i}=1}^{\mathrm{Q}}\sum_{\mathrm{j}=1}^{\mathrm{Q}}\gamma_{\mathrm{B_i}}^{\mathrm{S_0}}\gamma_{\mathrm{B_j}}^{\mathrm{S_0}}\mathrm{Cov}(\mathrm{B_i}, \mathrm{B_j}|\mathbf{X})\Big]$, together with $\mathrm{E}_{\mathrm{B|X}}(\mu_0^{\mathrm{S_k}})$, we then obtain an approximation of $\mathrm{V}_{\mathrm{B|X}}(\mu_0^{\mathrm{S_k}}) = \mathrm{E}_{\mathrm{B|X}}[(\mu_0^{\mathrm{S_k}})^2] - \mathrm{E}_{\mathrm{B|X}}(\mu_0^{\mathrm{S_k}})$. Given $\hat{\boldsymbol{\beta}}_{\mathrm{X}}, \hat{\boldsymbol{\gamma}}_{\mathrm{B}}^{\mathrm{S_0}}, \hat{\mathrm{E}}(\mathbf{B}|\mathbf{X})$, and $\hat{\gamma}_0^{\mathrm{S_k}}$, we can obtain $\hat{\boldsymbol{\gamma}}_{\mathrm{X}}^{\mathrm{S_k}} = \hat{\boldsymbol{\beta}}_{\mathrm{X}}(1 - \frac{\hat{\mathrm{V}}_{\mathrm{B|X}}(\mu_0^{\mathrm{S_k}})}{\hat{\mathrm{E}}_{\mathrm{B|X}}(\mu_0^{\mathrm{S_k}})[1-\hat{\mathrm{E}}_{\mathrm{B|X}}(\mu_0^{\mathrm{S_k}})]})^{-1} - \Big[\hat{\mathrm{E}}(\mathbf{B}^{\mathrm{T}}|\mathbf{X}+1_{\mathrm{p}}) - \hat{\mathrm{E}}(\mathbf{B}^{\mathrm{T}}|\mathbf{X})\Big]\hat{\boldsymbol{\gamma}}_{\mathrm{B}}^{\mathrm{S_0}}$.

Note that we estimate $\mathrm{E}(\mathbf{B}|\mathbf{X}; \boldsymbol{\theta}) = \mathrm{g}'^{-1}(\boldsymbol{\theta}\mathbf{X})$ and $\mathrm{Var}(\mathbf{B}|\mathbf{X}; \boldsymbol{\theta}) = \mathrm{g}'^{-1}(\boldsymbol{\theta}\mathbf{X})\Big[1 - \mathrm{g}'^{-1}(\boldsymbol{\theta}\mathbf{X})\Big]$ using the internal data by regressing each B on $\mathbf{X}$ with appropriate link function $\mathrm{g}'()$ based on the type of B, e.g., when B is continuous, linear regression and identity link is used; when B is binary, logistic regression and logit link is used. Given $\hat{\boldsymbol{\theta}}$, $\hat{\mathrm{E}}(\mathbf{B}|\mathbf{X}) = \hat{\boldsymbol{\theta}}^{\mathrm{T}}\mathrm{E}(\mathbf{X})$ is used.

C.2 Additional Simulation Results

In this section, we show the results of additional simulations to assess the performance of the proposed strategy for point estimates and variance estimation.

C.2.1 Continuous outcome Y (a supplement to Simulation I in Chapter 4)

Goal: To examine the proposed method when the outcome is continuous and the target model is linear regression.

Simulation setup: This simulation is the same as Simulation I in Chapter 4, except the generative outcome model now follows Gaussian distribution:

$$\begin{cases} \text{Internal:} & \mathrm{Y}|\mathbf{X},\mathbf{B} \sim \mathrm{N}(-1-\mathrm{X}_1-\mathrm{X}_2-\mathrm{B}_1-\mathrm{B}_2,1); \\ \text{External 1:} & \mathrm{Y}|\mathbf{X},\mathbf{B} \sim \mathrm{N}(1-\mathrm{X}_1-\mathrm{X}_2-\mathrm{B}_1-\mathrm{B}_2,1); \\ \text{External 2:} & \mathrm{Y}|\mathbf{X},\mathbf{B} \sim \mathrm{N}(3-\mathrm{X}_1-\mathrm{X}_2-\mathrm{B}_1-\mathrm{B}_2,1). \end{cases}$$

The target outcome model (model 2 in Chapter 4) is now a linear regression:

$$\mathrm{E}(\mathrm{Y}|\mathbf{X},\mathbf{B},\mathbf{S}) = \gamma_0 + \sum_{\mathrm{k}=1}^{2} \gamma_0^{\mathrm{S_k}} \mathrm{S_k} + \sum_{\mathrm{p}=1}^{2} \gamma_{\mathrm{X_p}} \mathrm{X_p} + \sum_{\mathrm{q}=1}^{2} \gamma_{\mathrm{B_q}} \mathrm{B_q},$$

Results: Figure C.1 shows similar pattern as those in Simulation I in Chapter 4, where the proposed method (red dotted curve) has the smallest bias among all for all covariates (Figure C.1a), largest precision gain compared with others (Figure C.1b), and the closest variance estimation to the Monte Carlo empirical variance (Figure C.1c).

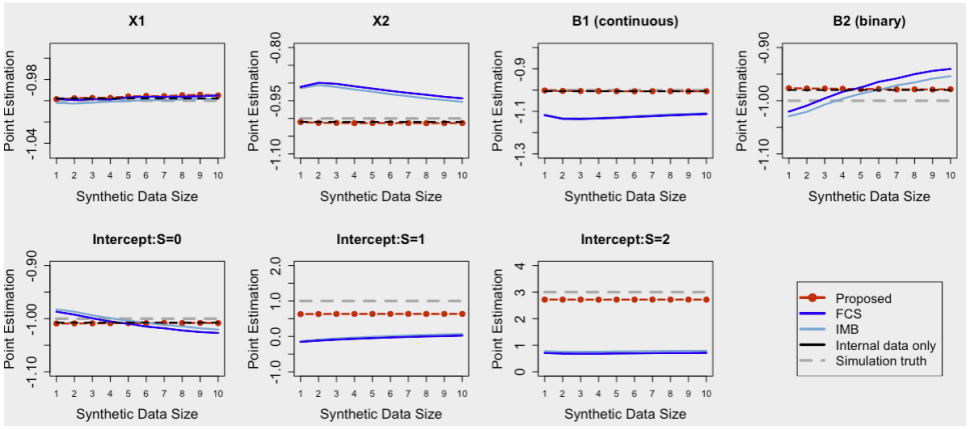
C.2.2 Smaller covariate effect (a modification to Simulation I in Chapter 4)

Goal: To assess our approach when the magnitude and the difference of covariate effects are small across different populations in the target outcome model.

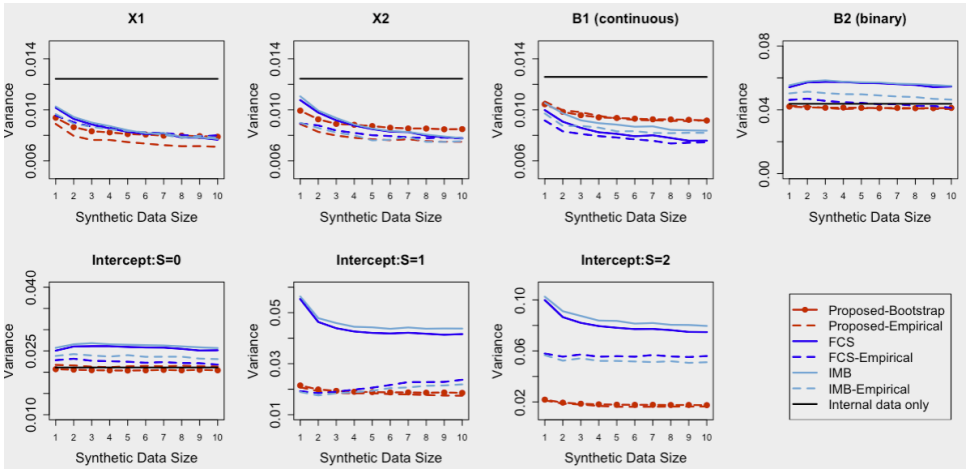
Simulation setup: This simulation is the same as Simulation I in Chapter 4, except the coefficient effect is now -0.5 instead of -1, and the intercept difference is smaller among populations:

$$\begin{cases} \text{Internal:} & \mathrm{logit}[\Pr(\mathrm{Y}=1|\mathbf{X},\mathbf{B})] = -1-0.5\big(\mathrm{X}_1+\mathrm{X}_2+\mathrm{B}_1+\mathrm{B}_2\big), \text{ prevalence=0.28;} \\ \text{External 1:} & \mathrm{logit}[\Pr(\mathrm{Y}=1|\mathbf{X},\mathbf{B})] = -0.5-0.5\big(\mathrm{X}_1+\mathrm{X}_2+\mathrm{B}_1+\mathrm{B}_2\big), \text{ prevalence=0.36;} \\ \text{External 2:} & \mathrm{logit}[\Pr(\mathrm{Y}=1|\mathbf{X},\mathbf{B})] = 0-0.5\big(\mathrm{X}_1+\mathrm{X}_2+\mathrm{B}_1+\mathrm{B}_2\big), \text{ prevalence=0.45.} \end{cases}$$

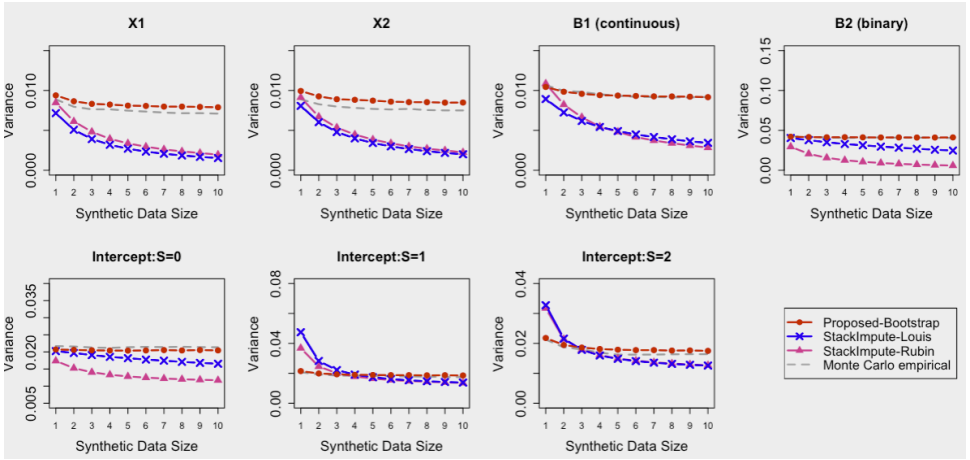
Results: Figure C.2a shows that compared with larger covariate effects in Simulation I in Chapter 4, when the X covariate effect is small, FCS and IMB have smaller bias in estimating X coefficients but still lack the ability to identify population-specific effects (i.e. intercepts of external populations). Similarly, Figure C.2b shows smaller bias of variance estimation. Note that the Rubin's rule variance estimator in Figure C.2b is too large (the pink curve) so that it falls outside of the range of the figure.



(a) Point estimates

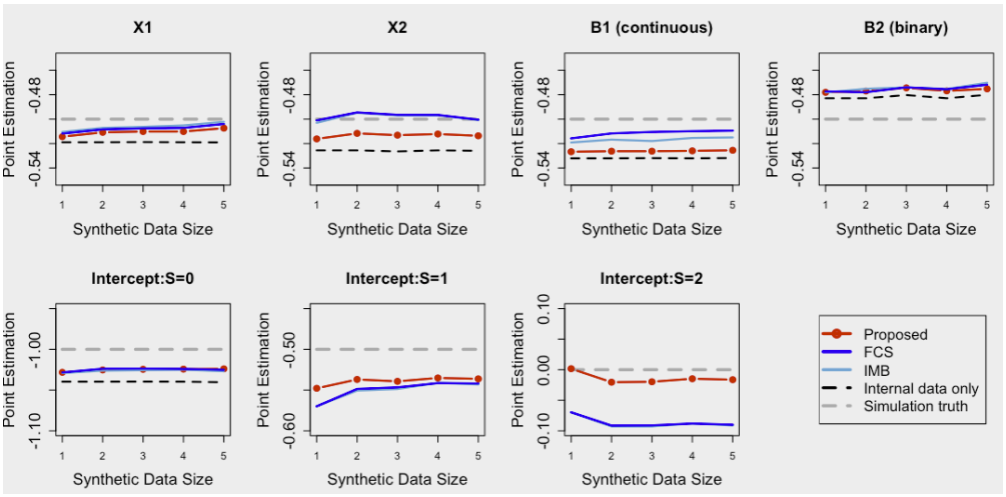


(b) Variance estimator vs. the empirical variance

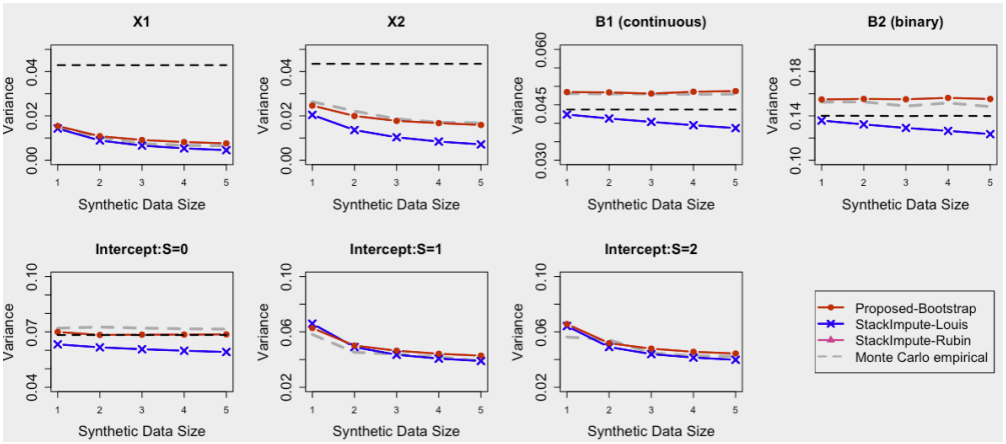


(c) Different variance estimators of the proposed method

Figure C.1: Results of Simulation C.2.1 over increasing synthetic data size (a) point estimates (b)variance estimation vs. Monte Carlo empirical variance (c) different variance estimators of the proposed method.



(a) Point estimates



(b) Different variance estimators of the proposed method

Figure C.2: Results of Simulation C.2.2 over increasing synthetic data size (a) point estimates (b) different variance estimators of the proposed method.

C.2.3 Different X covariate effects in the outcome model (a more flexible outcome model compared with Simulation I in Chapter 4)

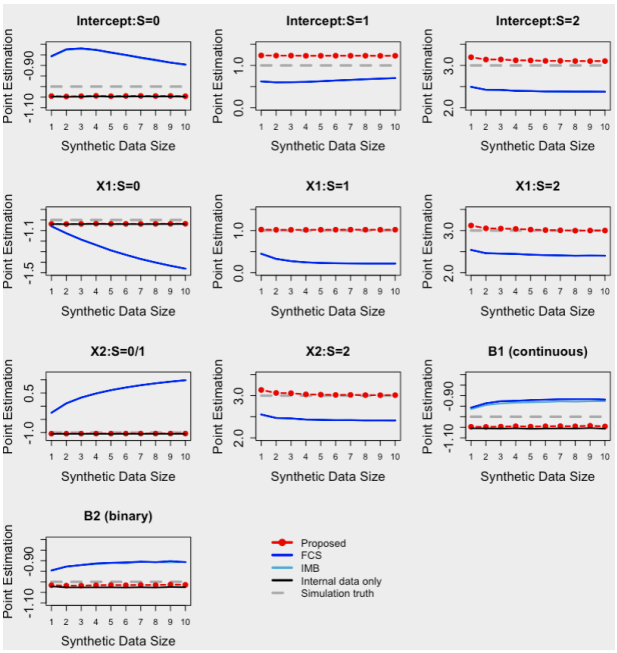
Goal: In Chapter 4, we only present the simulation results allowing the target model's intercept to differ across populations. In this simulation, we additionally show the performance of the proposed method when all possible X covariates coefficients are allowed to differ across populations (similar to model 1 or "different intercept and covariates" model in the real data example in Chapter 4).

Simulation setup: This simulation is the same as Simulation I in Chapter 4 except now that the

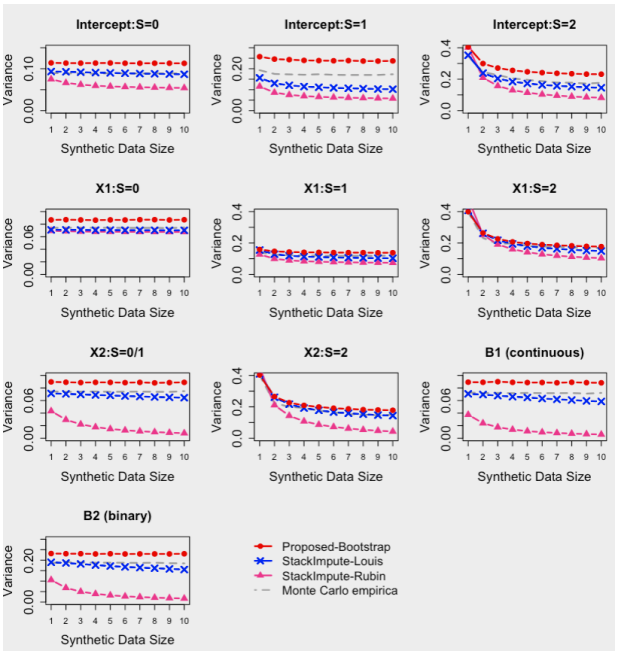
generative outcome models are as follows:

$$\begin{cases} \text{Internal:} & \text{logit}[\Pr(\mathrm{Y}=1|\mathbf{X},\mathbf{B})] = -1 - \mathrm{X}_1 - \mathrm{X}_2 - \mathrm{B}_1 - \mathrm{B}_2, \text{ prevelance= 0.3;} \\ \text{External 1:} & \text{logit}[\Pr(\mathrm{Y}=1|\mathbf{X},\mathbf{B})] = 1 + \mathrm{X}_1 - \mathrm{X}_2 - \mathrm{B}_1 - \mathrm{B}_2, \text{ prevelance= 0.58;} \\ \text{External 2:} & \text{logit}[\Pr(\mathrm{Y}=1|\mathbf{X},\mathbf{B})] = 3 + 3\mathrm{X}_1 + 3\mathrm{X}_2 - \mathrm{B}_1 - \mathrm{B}_2, \text{ prevelance= 0.70.} \end{cases}$$

Results: Similar to the results of Simulation I in Chapter 4, the results in Figure C.3 shows outstanding performance of the proposed method in both point estimates and variance estimation compared with others. For example, the proposed method has small bias less than 0.02 when estimating X_2 in population S=1 while the bias in FCS and IMB can go up to 0.78 (i.e. almost 40 times of the proposed method).



(a) Point estimates



(b) Different variance estimators of the proposed method

Figure C.3: Results of Simulation C.2.3 over increasing synthetic data size (a) point estimates (b) different variance estimators of the proposed method.

C.2.4 Violation of transportability assumption

Goal: To examine the proposed method when Assumption 2 ($\mathrm{X}_{\text{miss}}|\mathrm{X}_{\text{obs}}$ and $\mathrm{B}|\mathrm{X}$ are transportable between the internal and the external populations) is violated. We present two examples where the violation only causes ignorable bias in case 1 while it has larger impact in case 2.

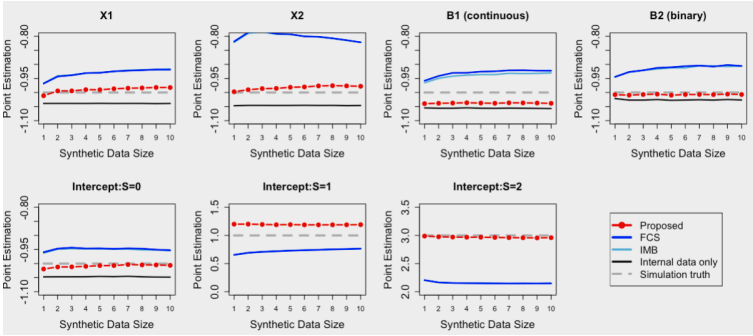
C.2.4.1 Case 1: different $\mathrm{B}|\mathrm{X}$ distribution in external population 2

Simulation setup: This simulation is the same as Simulation I in Chapter 4 except that now the external model 2 has different marginal B_1 distribution and different conditional distribution $\mathrm{B}_2|\mathrm{X}_1, \mathrm{X}_2, \mathrm{B}_1$:

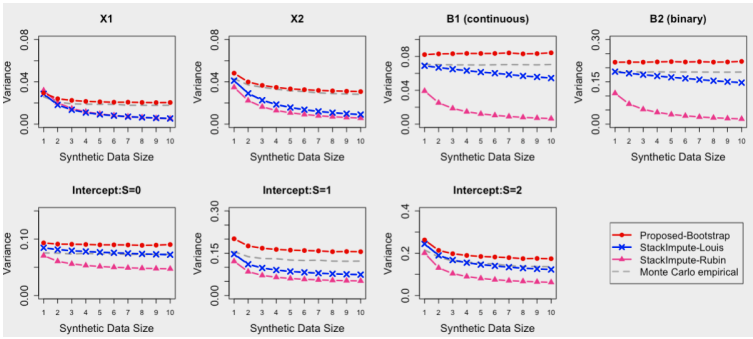
- B_1 has mean 1.5 and standard deviation 1.5 in external population 2 while in other populations B_1 has mean 0 and standard deviation 1;
- $\mathrm{B}_2|\mathbf{X}, \mathrm{B}_1 \sim \mathrm{Ber}\{[1 + \exp^{-1}(0.2\mathrm{X}_1 + 0.3\mathrm{X}_2 + 0.4\mathrm{B}_1)]\}$ in external population 2 while in other populations $\mathrm{B}_2|\mathbf{X}, \mathrm{B}_1 \sim \mathrm{Ber}\{[1 + \exp^{-1}(0.1\mathrm{X}_1 + 0.2\mathrm{X}_2 + 0.3\mathrm{B}_1)]\}$.

Note that both B_1 and B_2 are only observed in the internal study and multiple imputations are needed for them, where $\mathrm{B}_2|\mathbf{X}$ and $\mathrm{B}_2|\mathbf{X}, \mathrm{B}_1$ should be the same across populations according to Assumption 2.

Results: Figure C.4a indicates that the violation of transportability assumption in the proposed method has limited impact of point estimation with ignorable bias while Figure C.4b shows similar pattern of variance estimations as before.



(a) Point estimates



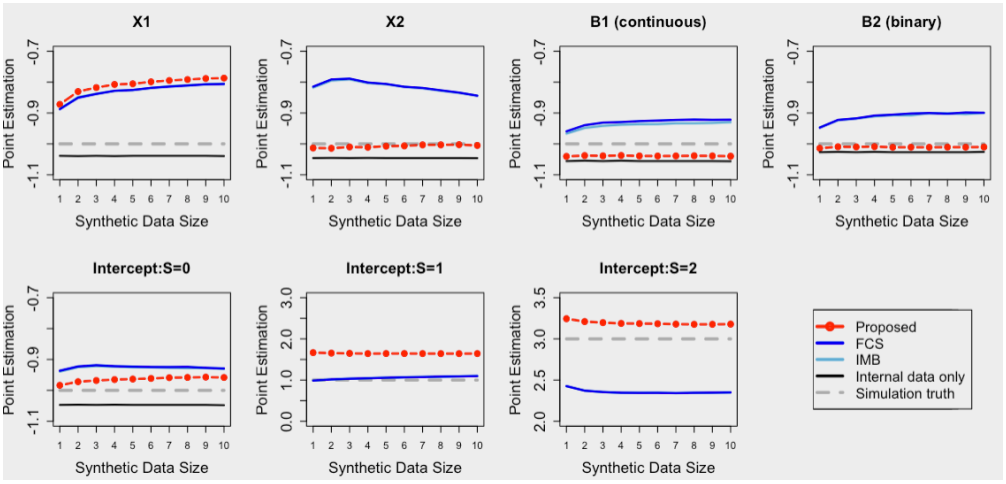
(b) Different variance estimators of the proposed method

Figure C.4: Results of Simulation C.2.4.1 over increasing synthetic data size (a) point estimates (b) different variance estimators of the proposed method.

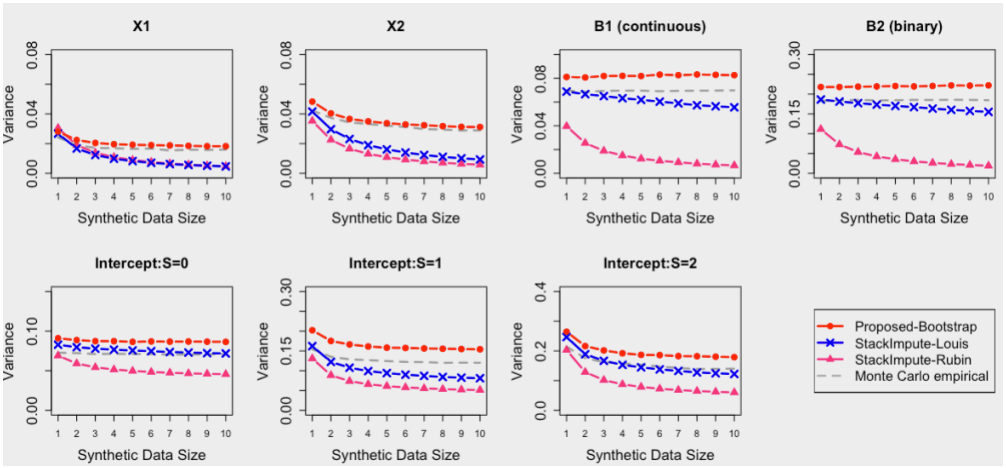
C.2.4.2 Case 2: different marginal X_1 distribution in external populations

Simulation setup: This simulation is the same as Simulation I in Chapter 4 except now that in the external studies, $X_1 \sim N(1, 1.5)$ while in the internal study $X_1 \sim N(0, 1)$. This will lead to different distribution conditional on X_1 and thus violates Assumption 2.

Results: Figure C.5a shows that such violation leads to some bias of estimated coefficient X_1, i.e., 0.2 absolute bias. Besides that, the proposed method has nearly unbiased point estimates for other parameter (i.e. up to 0.014 absolute bias) while the bias in FCS and IMB can be up to 15 times the bias of the proposed method. Similarly, Figure C.5b shown unbiased variance estimation of the proposed bootstrap estimator.



(a) Point estimates



(b) Different variance estimators of the proposed method

Figure C.5: Results of Simulation C.2.4.2 over increasing synthetic data size (a) point estimates (b) different variance estimators of the proposed method.

BIBLIOGRAPHY

K. Viele, S. Berry, B. Neuenschwander, B. Amzal, F. Chen, N. Enas, B. Hobbs, J. G. Ibrahim, N. Kinnersley, S. Lindborg, S. Micallef, S. Roychoudhury, and L. Thompsonl. Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics*, 13:41–54, 2014.

D. Dejardin, P. Delmar, C. Warne, K. Patel, J. van Rosmalen, and E. Lesaffre. Use of a historical control group in a noninferiority trial assessing a new antibacterial treatment: A case study and discussion of practical implementation aspects. *Pharmaceutical Statistics*, 17:169–181, 2018.

X. H. Li and Y. Song. Target population statistical inference with data integration across multiple sources—an approach to mitigate information shortage in rare disease clinical trials. *Statistics in Biopharmaceutical Research*, 12:322–333, 2020.

C. Bycroft. Integrated household surveys: a survey vehicles approach. Technical report, Wellington, 1 2011.

S. Yang and J. K. Kim. Statistical data integration in survey sampling: a review. *Japanese Journal of Statistics and Data Science*, 3:625–650, 2020.

S. Yang and P. Ding. Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*, 3:1540–1554, 2020a.

M. H. Gail, L. A. Brinton, D. P. Byar, D. K. Corle, S. B. Green, C. Schairer, and J. J. Mulvihill. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute*, 81:1879–1886, 1989.

I. M. Thompson, D. P. Ankerst, C. Chi, P. J. Goodman, C. M. Tangen, M. S. Lucia, Z. Feng, H. L. Parnes, and C. A. Coltman. Assessing prostate cancer risk: Results from the prostate cancer prevention trial. *European Urology*, 98:529–534, 2006.

M. J. Roobol, H. A. van Vugt, S. Loeb, X. Zhu, M. Bul, C. H. Bangma, A. G. van Leenders, E. W. Steyerberg, and F. H. Schröder. Prediction of prostate cancer risk: the role of prostate volume and digital rectal examination in the ERSPC risk calculators. *European Urology*, 61:577–583, 2012.

C. Stephan, B. Vogel, H. Cammann, M. Lein, V. Klevecka, P. Sinha, G. Kristiansen, D. Schnorr, K. Jung, and S. A. Loening. An artificial neural network as a tool in risk evaluation of prostate cancer. indication for biopsy with the psa range of 2-20 microg/l. *Der Urologe, Ausgabe A*, 42: 1221–1229, 2003. doi: https://doi.org/10.1007/s00120-003-0322-7.

A. Osareh and B. Shadgar. Machine learning techniques to diagnose breast cancer. *2010 5th International Symposium on Health Informatics and Bioinformatics*, pp:114–120, 2010. doi: doi:10.1109/HIBIT.2010.5478895.

H. Estiri, Z.H. Strasser, J.G. Klann, P. Naseri, K. B. Wagholikar, and S. N. Murphy. Predicting covid-19 mortality with electronic medical records. *npj Digit. Med*, 4:15, 2021. doi: https://doi-org.proxy.lib.umich.edu/10.1038/s41746-021-00383-x.

G. W. Imbens and T. Lancaster. Combining micro and macro data in microeconometric models. *The Review of Economic Studies*, 61:655–680, 1994.

S. Grill, M. Fallah, R. J. Leach, I. M. Thompson, K. Hemminki, and D. P. Ankerst. A simple-to-use method incorporating genomic markers into prostate cancer risk prediction tools facilitated future validation. *Journal of Clinical Epidemiology*, 68:563–573, 2015.

E. Bareinboim and J. Pearl. A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference*, 1:107–134, 2013.

J. Qin. Combining parametric and empirical likelihoods. *Biometrika*, 87:484–490, 2000.

P. Han and J. F. Lawless. Empirical likelihood estimation using auxiliary summary information with different covariate distribution. *Statistics Sinica*, 29:1321–1342, 2019.

N. Chatterjee, Y.-H. Chen, P. Maas, and R. J. Carroll. Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association*, 111:107–117, 2016.

W. Cheng, J. M. G. Taylor, P. S. Vokonas, S. K. Park, and B. Mukherjee. Improving estimation and prediction in linear regression incorporating external information from an established reduced model. *Statistics in Medicine*, 37:1515–1530, 2018.

W. Cheng, J. M. G. Taylor, T. Tomlins, and B. Mukherjee. Informing a risk prediction model for binary outcomes with external coefficient information. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68:121–139, 2019.

S. Grill, D. P. Ankerst, M. H. Gail, N. Chatterjee, and R. M. Pfeiffer. Comparison of approaches for incorporating new information into existing risk prediction models. *Statistics in Medicine*, 36:1134–1156, 2017.

J. P. Estes, B. Mukherjee, and J. M. G. Taylor. Empirical Bayes estimation and prediction using summary-level information from external big data sources adjusting for violations of transportability. *Statistics in Biosciences*, 10:568–586, 2017.

P. Kundu, R. Tang, and N. Chatterjee. Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika*, 106:567–585, 2019.

Z. Chen, J. Ning, Y. Shen, and J. Qin. Combining primary cohort data with external aggregate information without assuming comparability. *Biometrics*, 2020. doi: 10.1111/biom.13356.

S. Yang and P. Ding. Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*, 115:1540–1554, 2020b.

T. E. Raghunathan, J. P. Reiter, and D. B. Rubin. Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19:1, 2003.

J. P. Reiter and S. K. Kinney. Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary. *Journal of Official Statistics*, 28:583–n/a, 2012.

J. P. Reiter. Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, 18:531, 2002.

D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons Inc., New York, 1987.

S. Van Buuren and C. G. M. Oudshoom. MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45, 2011. doi: 10.18637/jss.v045.i03.

D. Xu, M. J. Daniels, and A. G. Winterstein. Sequential BART for imputation of missing covariates. *Biostatistics*, 17:589–602, 2016.

S. A. Tomlins, J. R. Day, R. J. Lonigro, D. H. Hovelson, J. Siddiqui, L. P. Kunju, R. L. Dunn, S. Meyer, P. Hodge, J. Groskopf, J. T. Wei, and A. M. Chinnaiyan. Urine TMPRSS2:ERG plus PCA3 for individualized prostate cancer risk assessment. *European Urology*, 70:45–53, 2015.

H. Zhang, L. Deng, M. Schiffman, J. Qin, and K. Yu. Generalized integration model for improved statistical inference by leveraging external summary data. *Biometrika*, 107:689–703, 2020.

P. Goos and B. Jones. *Optimal Design of Experiments: A Case Study Approach*. Wiley, New Jersey, 2011.

M. Truong, B. Yang, and D. F. Jarrard. Toward the detection of prostate cancer in urine: a critical analysis. *Journal of Urology*, 189:422–429, 2013.

A. M. Bohnen, F. P. Groeneveld, and J. L. H. R. Bosch. Serum prostate-specific antigen as a predictor of prostate volume in the community: the krimpen study. *European Urology*, 51: 1645–1653, 2007.

A. J. Simpkin, J. L. Donovan, K. Tilling, J. Athene Lane, R. M. Martin, P.C. Albertsen, A. Bill-Axelson, H. Ballentine Carter, J. L. Bosch, L. Ferrucci, F. C. Hamdy, L. Holmberg, E. Jeffrey Metter, D. E. Neal, C. C. Parker, and C. Metcalfe. Prostate-specific antigen patterns in us and european populations: comparison of six diverse cohorts. *BJU International*, 118:911–918, 2016.

R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5:197–227, 1990.

Y. Freund and R. R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

L. Breiman. Bagging predictors. *Random forests*, 45:5–32, 2001.

M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6:25, 2007.

F. Wang, L. Wang, and P. X.-K. Song. Quadratic inference function approach to merging longitudinal studies: validation and joint estimation. *Biometrika*, 99(3):755–762, 09 2012a. ISSN 0006-3444. doi: 10.1093/biomet/ass021.

J. Antonelli, C. Zigler, and F. Dominici. Guided Bayesian imputation to adjust for confounding when combining heterogeneous data sources in comparative effectiveness research. *Biostatistics*, 18:553–568, 2017.

C. Wang, G. Parmigiani, and F. Dominici. Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics*, 68:661–671, 2012b.

L. J. Beesley and J. M. G. Taylor. A stacked approach for chained equations multiple imputation incorporating the substantive model. *Biometrics*, 2020. doi: 10.1111/biom.13372.

L. J. Beesley and J. M. G. Taylor. Accounting for not-at-random missingness through imputation stacking. *ArXiv*, 2021. doi: 2101.07954.

J. Neuhaus and N. Jewell. A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika*, 80:807–815, 1993.

S. Van Buuren, J. P. L. Brand, C. G. M. Groothuis-Oudshoorn, and D. B. Rubin. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76: 1049–1064, 2006.

F. Li, M. Baccini, F. Mealli, E. R. Zell, C. E. Frangakis, and D. B. Rubin. Multiple imputation by ordered monotone blocks with application to the anthrax vaccine research program. *Journal of Computational and Graphical Statistics*, 23(3):877–892, 2014. doi: 10.1080/10618600.2013.826583.

R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley and Sons, Inc, Hoboken, NJ, 2nd edition, 2002.

J. Y. Dai, C. Kooperberg, M. Leblanc, and R. L. Prentice. Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. *Biometrika*, 99:929–944, 2012.

C. Huang and J. Qin. A unified approach for synthesizing population-level covariate effect information in semiparametric estimation with survival data. *Statistics in Medicine*, 39:1573–1590, 2020.

S. Rassler. Data fusion: identification problems, validity, and multiple imputation. *Statistica Sinica*, 33:153–171, 2004.

J. P. Reiter. Bayesian finite population imputation for data fusion. *Statistica Sinica*, 22:795–811, 2012.

C. Gourieroux and A. Monfort. On the problem of missing data in linear models. *Review of Economic Studies*, 48:579–586, 1981.

G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University, Baltimore, MD, 3rd edition, 1996.